Interpolated retrieval effects on list isolation: Individual differences in working memory capacity

By: Christopher N. Wahlheim, Timothy R. Alexander, and Michael J. Kane

Wahlheim, C.N., Alexander, T.R., & Kane, M.J. (2019). Interpolated retrieval effects on list isolation: Individual differences in working memory capacity. *Memory & Cognition*, 47, 619-642.

This is a post-peer-review, pre-copyedit version of an article published in *Memory & Cognition*. The final authenticated version is available online at: http://dx.doi.org/10.3758/s13421-019-00893-w

Abstract:

We examined the effects of interpolated retrieval from long-term memory (LTM) and short-term memory (STM) on list isolation in dual-list free recall and whether individual differences in working memory capacity (WMC) moderated those effects. Ninety-seven subjects completed study–test trials that included two study lists separated by either an exemplar generation task (LTM retrieval) or a two-back task (STM retrieval). Subjects then completed an externalized free recall task that allowed for the examination of response accessibility and monitoring. Individual differences in WMC were assessed using three complex span tasks: operation span, reading span, and rotation span. Correct recall and intratrial intrusion summary scores showed no effect of interpolated retrieval on either response accessibility or monitoring. However, serial-position curves for correct recall of List 1 showed larger primacy in the two-back than in the exemplar generation task for high-WMC subjects. We interpreted these results from a context change perspective, as showing that interpolated LTM retrieval accelerated context change for subjects who processed the context more effectively. We consider the implications of these findings for models of memory.

Keywords: Context change | Control processes | Free recall | Interference | Working memory

Article:

> Because consciousness is equated with the short-term store and because control processes are centered in and act through it, the short-term store is considered a working memory: a system in which decisions are made, problems solved and information flow is directed. . . . So much information is contained in the long-term store that the major problem is finding access to some small subset of the information that contains the desired image, just as one must find a particular book in a library before it can be scanned for the desired information. (Atkinson & Shiffrin, 1971, p. 83)

Successful retrieval of episodic memories depends on the ability to isolate specific past experiences while avoiding interference from competing memories. In list-learning experiments, this can take the form of recalling from a target list while avoiding intrusions from another list. List isolation depends on the cognitive activities that occur between lists and how subjects can exert control over such activities. Our goal in the present experiment is to examine how the retrieval processes engaged between lists affects list isolation and whether this depends on control abilities. Specifically, we examine how list isolation differs when interpolated tasks require retrieval from long-term memory (LTM) or short-term memory (STM), and whether such effects interact with individual differences in working memory capacity (WMC). We adopt the theoretical orientation that interpolated retrieval has its effects by inducing context change. To explain individual differences, we consider theoretical perspectives on the relationship between control processes and context utilization from dual-store and WMC models.

Retrieval-induced list isolation

A central theme in the list isolation literature is that interpolated retrieval differentiates lists by accelerating internal context change (e.g., Divis & Benjamin, 2014; Jang & Huber, 2008; Sahakyan & Hendricks, 2012). In list-learning experiments, internal context refers to the mental states that become associated with episodic representations of lists and their constituent items. Some models propose that list contexts remain constant across time (e.g., DeLosh & McDaniel, 1996; Rohrer, 1996; Rohrer & Wixted, 1994), whereas others assume that context alternates between active and inactive states, producing random fluctuation (e.g., Estes, 1955; Mensink & Raaijmakers, 1988). For the latter, the alternation produces contextual drift across items that is independent of the items themselves. The Temporal Context Model (Howard & Kahana, 2002) has specified a contextual drift mechanism by proposing that individual items can drive context change. According to the model, items themselves elicit retrieval of earlier contexts, which updates the current state of context (see also Lohnas, Polyn, & Kahana, 2015; Polyn, Norman, & Kahana, 2009). The present experiment tests this assumption by examining the effects of interpolated retrieval on list isolation.

Interpolated retrieval effects on list isolation were initially shown by Shiffrin (1970) using the "list-before-last" paradigm. Subjects studied word lists of varying lengths and were asked to recall the list prior to the last list. Intervening list length did not affect target list recall. This was interpreted as showing that interpolated retrieval isolated target lists, which aided in later context reinstatement during recall (see also Klein, Shiffrin, & Criss, 2007). Subsequent work has confirmed that list isolation depends on retrieval occurring between lists (e.g., Jang & Huber, 2008; Unsworth, Spillers, & Brewer, 2012; Ward & Tan, 2004). Most relevant to the present study, Jang and Huber (2008) found list isolation effects when subjects retrieved from LTM between lists, in tasks such as category exemplar generation, but not when they retrieved from short-term memory (STM), as in a two-back task. These results suggested that LTM retrieval was necessary to produce the context change that led to list isolation.

Similar conclusions have been drawn from other studies. For example, Sahakyan and Hendricks (2012) used a list-before-last paradigm to examine whether list isolation effects varied with interpolated-retrieval difficulty. Subjects studied three lists of words, completed an interpolated

task between Lists 2 and 3, and took a final recall test on List 2. The interpolated task required subjects to either recall items from List 1, or to complete math problems in a control condition. Interpolated retrieval led to list isolation, as indicated by a reduction in List 2 recall and List 3 intrusions relative to the control condition. These results were taken as evidence that retrieval accelerated context change. Similarly, Divis and Benjamin (2014) found that interpolated retrieval from LTM reduced proactive interference in a multiple-list learning paradigm. Subjects studied five lists of words with an interpolated task between each. One trial included an interpolated category exemplar generation task (i.e., LTM retrieval) and another trial included an interpolated counting-backward task (i.e., distractor control). Fifth-list recall performance was higher when the interpolated task was category exemplar generation rather than a control task, whereas the reverse was true for first-list recall. Prior-list intrusions on fifth-list recall were also lower in the interpolated category exemplar generation condition than in the distractor-control condition. These findings converged with earlier results in suggesting that interpolated LTM retrieval led to list isolation by accelerating context change.

Despite these demonstrations that retrieval from LTM is necessary to produce list isolation effects, other work has found no differences in list isolation between interpolated LTM and STM retrieval. For example, in multiple-list learning, Pastötter, Schicker, Niedernhuber, and Bäuml (2011) found that final-list recall performance did not differ when the interpolated task was category exemplar generation (LTM retrieval) or a two-back task (STM retrieval). Contrary to the findings above, these results suggest that retrieval from LTM may be sufficient but not necessary to drive list isolation. However, Pastötter et al. also found fewer prior list intrusions during final-list recall when interpolated retrieval was from LTM rather than from STM, which provides some evidence that interpolated LTM retrieval reduced proactive interference. These mixed effects suggest that it may useful to further examine the conditions and subject characteristics that determine when interpolated LTM retrieval will drive list isolation.

List isolation and individual differences in WMC

According to a contextual differentiation account (e.g., Sahakyan & Kelley, 2002), tasks external to the subject can influence the rate at which internal contextual elements change. However, the effects of task manipulations on context change may also depend on internal factors, such as individual differences in control abilities. Indeed, subject-by-condition interactions arise in related work on list-method directed forgetting. Delaney and Sahakyan (2007) showed that WMC predicted differences in directed forgetting costs and benefits. Subjects studied two lists of words, took free recall tests for each list, and completed a WMC task. Some subjects were instructed to forget List 1 prior to studying List 2 (*forget* condition), while others were instructed to remember List 1 (*remember* condition). In the forget condition, higher-WMC subjects showed lower recall of List 1 and higher recall of List 2 relative to lower-WMC subjects (for similar results, see Aslan, Zellner, & Bäuml, 2010; Soriano & Bajo, 2007). These results suggested that higher-WMC subjects more effectively (or actively) used control processes to change mental context than did lower-WMC subjects.

Related to this, individual differences in WMC moderate list isolation even without an interpolated-task manipulation. Unsworth (2009) had subjects complete a version of the list-before-last recall task along with three measures of WMC. In the recall task, subjects completed

several trials in which they studied two lists of words and were postcued to recall from only one of those lists. Cluster analyses assessed associations between WMC and free recall measures. Higher-WMC subjects clustered into groups that: recalled many correct items and produced few intrusions, or recalled fewer correct items but also showed average intrusions. In contrast, lower-WMC subjects clustered into groups that: showed comparable rates of correct recall and intrusion production, or average recall with high rates of intrusions. These results were interpreted as showing that higher-WMC subjects more effectively focused their memory search to target lists and monitored the source of retrievals. Although these findings do not bear directly on the effects of retrieval-induced context change, it has been hypothesized that such individual differences reflect differences in context processing, for example, in storing contextual information at encoding or using temporal context cues to guide or delimit subsequent retrieval (Sahakyan, Abushanab, Smith, & Gray, 2014; Spillers & Unsworth, 2011; Unsworth, 2007, 2016; Unsworth, Brewer, & Spillers, 2011; Unsworth & Engle, 2007; Unsworth & Spillers, 2010). If so, these results could be interpreted as showing that the rate of context change across phases in the experiment differed on the basis of subjects' WMC. Together, these studies suggest that the effects of interpolated LTM retrieval may also depend on WMC.

The present experiment

We examined whether the effects of interpolated retrieval from LTM and STM interact with WMC, using a dual-list free recall paradigm in which subjects studied two lists of words and were postcued to recall from only one list (cf. Unsworth, 2009; Unsworth, Brewer, & Spillers, 2013; Wahlheim, Ball, & Richmond, 2017; Wahlheim & Huff, 2015; Wahlheim, Richmond, Huff, & Dobbins, 2016). Between each study list, subjects completed category exemplar generation or two-back tasks on separate trials. We considered the primary difference between these tasks to be that category-exemplar generation relied more on LTM retrieval than did two-back judgments. Of course, these tasks differ in other ways, such as their reliance on semantic information and perhaps their overall difficulty, but for consistency with earlier studies (e.g., Jang & Huber, 2008), we highlight differences between LTM and STM retrieval.

Another key feature of the present paradigm is that we used an externalized free recall (EFR) procedure (Bousfield & Rosner, 1970; Roediger & Payne, 1985) to assess both accessibility and monitoring of responses. We took this approach to more precisely characterize the effects of interpolated retrieval, and thus to inform models of free recall, as described below. In the typical EFR procedure, subjects are asked to recall from a specific source but to output *all* responses while doing so. In recent versions of this procedure, subjects must indicate which of their responses reflected source errors after each retrieval (e.g., Kahana, Dolan, Sauder, & Wingfield, 2005; Unsworth & Brewer, 2010). In our variant of EFR (which bears some similarity to modified–modified free recall procedures from the interference literature; e.g., Keppel, Postman, & Zavortink, 1967), subjects were instructed to recall from a target list, report all responses that came to mind while doing so, and indicate whether each response was correct (from the target list) or incorrect (not from the target list). After subjects completed the free recall task, we assessed WMC using three complex span tasks.

If we assume, on the basis of previous context-based accounts, that repeated retrieval from LTM results in greater context updating than does repeated retrieval from STM (e.g., Howard &

Kahana, 2002; Jang & Huber, 2008), then we should observe costs and benefits of interpolated retrieval like those of directed forgetting. That is, interpolated retrieval from LTM should lead to lower List 1 recall and higher List 2 recall than should interpolated retrieval from STM, and the reverse should be true for intrusions. In addition, if we assume that WMC allows, in part, the effective monitoring of retrieved context (e.g., Unsworth, 2009), then higher-WMC subjects should show better monitoring.

To formulate hypotheses about the relationship between WMC and interpolated-retrieval effects on list isolation, we consider two major theoretical perspectives on memory control processes and their predictions for the primary outcome measures in the present experiment. Note that we limit our primary discussion of model predictions to recall and intrusion summary scores. We leave more nuanced predictions about conditional recall measures for the Resultsand Discussion sections, in which we distinguish a priori predictions from more exploratory analyses.

Model predictions

Dual-store models. The original dual-store model by Atkinson and Shiffrin (1968) proposes an architecture of memory that includes permanent structural components along with transient control processes that govern encoding, rehearsal, and retrieval operations. Of particular relevance here, control processes were proposed to select encoding strategies and to search and retrieve information from LTM. The theoretical control mechanism has had a long-standing influence on modern theories of a variety of episodic memory phenomena (e.g., Malmberg & Shiffrin, 2005; Raaijmakers & Shiffrin, 1980; Shiffrin & Steyvers, 1997). Lehman and Malmberg (2009), for example, proposed a buffer model that incorporates ideas from dual-store models to account for the effects of context change hypothesized to occur in directed forgetting. Most relevant to the present study, Lehman and Malmberg (2013) extended the model to explain serial-position effects and first-recall probabilities in free recall.

The buffer model makes three key assumptions that are relevant to the issues addressed in the present study. First, it assumes that a capacity-limited rehearsal buffer constrains the number of items that can be simultaneously rehearsed and the extent to which multiple items can be associated with the study context. Second, the model specifies a compartmentalization process that intentionally removes items from the buffer when those items are no longer needed for task completion. These complementary processes of rehearsal and compartmentalization control the contents of consciousness. Third, the model also assumes that a control process serves to monitor output decisions. Consequently, predictions about interactive effects of interpolated retrieval and WMC can be derived from this model.

If higher-WMC subjects have superior rehearsal and/or compartmentalization processes, then the predicted costs and benefits of interpolated LTM retrieval should be greater for them than for lower-WMC subjects. This would occur if higher-WMC subjects can more effectively use the compartmentalization process to drop items from consciousness during interpolated LTM retrieval and maintain items in consciousness during interpolated STM retrieval relative to lower-WMC subjects. This would create a greater difference in the accelerated context change conferred by interpolated LTM retrieval for higher- than for lower-WMC subjects. Furthermore,

if higher-WMC subjects can more effectively deploy control processes in the service of monitoring decisions than can lower-WMC subjects, then higher-WMC subjects should reject proportionally more intrusions that come to mind.

WMC theory. Predictions in the present experiment can also be derived from a theoretical approach to WMC functions proposed by Unsworth and Engle (2007) and further developed by Unsworth and colleagues (e.g., Unsworth et al., 2012). The WMC approach was also inspired by the suggestions (e.g., Atkinson & Shiffrin, 1968; Baddeley & Hitch, 1974) that memory-control processes guide cognitive operations across tasks. Unsworth and Engle (2007) originally suggested that WMC functions to (1) maintain active representations that include contextual information (e.g., Miyake & Shah, 1999) and (2) reinstate context cues to constrain memory search (e.g., Raaijmakers & Shiffrin, 1980). More recently, Unsworth and Brewer (2010) showed that WMC also contributes to the monitoring of retrievals from LTM.

The WMC functions proposed by Unsworth and colleagues have been supported by a variety of studies. For example, regarding active (goal) maintenance processes, higher-WMC subjects are better than lower-WMC subjects at avoiding attentional capture in antisaccade tasks (e.g., Kane, Bleckley, Conway, & Engle, 2001; Meier, Smeekens, Siliva, Kwapil, & Kane, 2018) and word reading errors in mostly congruent Stroop tasks (e.g., Kane & Engle, 2003; Meier & Kane, 2013), both of which require keeping novel goals, or task sets, accessible enough to control ongoing behavior and avoid habitual responding. The context reinstatement function has been demonstrated in free recall, where higher-WMC subjects recall more correct responses and fewer intrusions than do lower-WMC subjects (e.g., Unsworth & Engle, 2007). Additional evidence for this function of WMC has also been shown in free recall latencies, as higher-WMC subjects recall from target lists more rapidly than do lower-WMC subjects (e.g., Miller & Unsworth, 2018). The assumption here is that faster recall reflects more precise reinstatement of context, which leads to more correct items being recovered during LTM search. Finally, the monitoring function has been shown in EFR, as higher-WMC subjects appropriately classify more correct recalls and reject more intrusions than do lower-WMC subjects (e.g., Unsworth & Brewer, 2010).

If higher-WMC subjects maintain task sets and reinstate context more effectively than do lower-WMC subjects, then higher-WMC subjects should more effectively retrieve the contexts associated with items during interpolated LTM retrieval and maintain the ongoing task context during interpolated STM retrieval, relative to lower-WMC subjects. As in the buffer model, this should lead to more rapid acceleration of context change during interpolated LTM retrieval for higher- than for lower-WMC subjects. Furthermore, this hypothesized acceleration of context change for higher-WMC subjects leads to the prediction that greater context shifts during interpolated LTM retrieval will produce greater costs on the accessibility of earlier information and a greater match between the contexts of later information and the test phase for higher- relative to lower-WMC subjects. Combined with the assumption that higher-WMC subjects also monitor contextual information more effectively than do lower-WMC subjects, the WMC theory predicts that higher-WMC subjects should produce and appropriately classify more correct recalls and reject a greater proportion of intrusions than should lower-WMC subjects.

Thus, the dual-store buffer model and WMC theory make similar predictions regarding the interactive effects of WMC and interpolated LTM retrieval, but the mechanisms proposed to underlie these effects differ. The dual-store buffer model emphasizes that an active control process is used to maintain or discard items from consciousness, which determines the strength of associations that are formed between items and context. WMC theory proposes that an active control process serves to maintain task goals, which determines the contents of consciousness. However, WMC theory also specifies how differences in this control ability affect context reinstatement at retrieval. We contrast these models further when considering the implications of the present results for those perspectives in the Discussion.

Method

The full stimulus sets used in the present experiments, anonymized data files, coded data, and analysis script are available via the Open Science Framework: https://osf.io/fys48/. The research reported here was approved by the Institutional Review Board of the University of North Carolina at Greensboro (UNCG). We report how we determined our sample size, all data exclusions, all manipulations, and all measures below (Simmons, Nelson, & Simonsohn, 2012).

Subjects

We tested a total of 100 undergraduates at UNCG, who received partial credit toward a course requirement. Our goal was to test at least 96 subjects in a single semester, with our stopping rule for data collection being the end of the semester. According to G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009), with $N = 96$, we would have 85% power ($\alpha = .05$, two-tailed) to detect correlations of .30 between WMC and our dependent measures (and this sample size was divisible by our counterbalancing scheme of eight). It would take substantially larger samples than we were able to study for this initial investigation to detect weaker WMC correlations (e.g., $N = 193$ for 80% power to detect $\rho = .20$) or to allow a relatively stable effect-size estimate of any WMC correlations detected (Schönbrodt & Perugini, 2013). We will thus interpret WMC-related effect sizes cautiously throughout.

We succeeded in collecting data from the minimum number of planned subjects within the prespecified period (Simmons et al., 2012); testing four extra subjects offset our exclusion of three subjects' data from the free recall analyses due to failure to comprehend task instructions (two subjects) and to a program error (one subject). This resulted in a final sample of 97 subjects (60 females, 37 males) 18–25 years of age ($M = 18.76$, $SD = 1.40$); for the WMC-related analyses, however, we included only 95 subjects, because one subject did not complete the WMC tasks and one subject's WMC data were lost.

Design

We used a 2 (Trial Type: List 1 vs. List 2) × 2 (Interpolated Task: generation vs. two-back) within-subjects factorial design.

Materials

Here we describe the materials for the free recall and interpolated-retrieval tasks. We describe the complex span measures in the Procedure section.

Free recall. The stimuli included two sets of words taken from the MRC Psycholinguistic Database (Coltheart, 1981). One set served as practice items, and the other as critical items. The practice set consisted of 20 concrete nouns, four to nine letters in length (M = 5.33, SD = 1.49), with concreteness ratings of 525–624 (M = 572.35, SD = 25.73, scale = 100–700), and Hyperspace Analog to Language (HAL) log frequency (Lund & Burgess, 1996) counts of 6.54–12.25 (M = 9.59, SD = 1.46). The practice phase consisted of two study–test trials; each study list in the practice trial contained five words. We distributed four groups of five words matched on length, concreteness, and frequency across study lists. The lists remained constant across formats.

The critical item set consisted of 320 concrete nouns, four to nine letters in length (M = 5.43, SD = 1.37), with concreteness ratings of 502–670 (M = 578.87, SD = 30.53, scale = 100–700) and HAL log frequency counts of 6.94–12.60 (M = 9.63, SD = 1.12). The actual experiment consisted of 16 study–test trials, including four blocks of four trials each. Each block included one trial from each combination of the trial-type and interpolated-task conditions. That is, each block contained (a) two trials with an interpolated generation task, with subjects being postcued to recall from List 1 on one trial and List 2 on the other trial, and (b) two trials with an interpolated two-back task, with subjects being postcued to recall from List 1 on one trial and List 2 on the other. Each of these trials included two study lists with ten words each (4 blocks × 4 trials × 20 words per trial = 320 words total). We counterbalanced items across conditions by dividing the 320-word set into 32 groups of ten words that were matched on length, concreteness, and frequency. We then clustered the groups into four larger ensembles, each consisting of eight groups. We assigned each ensemble to one of the four trial blocks, and each group within each ensemble to a different study list. We then rotated the groups within the ensembles through lists and conditions, but fixed the assignment of ensembles to blocks across experiments formats. This resulted in words being presented equally often in each within-subjects condition, but not in each trial block. This scheme produced eight formats.

Interpolated retrieval tasks. The materials for the category exemplar generation task were nine category labels and 15 exemplar fragments from each of nine categories derived from the Van Overschelde, Rawson, and Dunlosky (2004) norms (for the complete material set, see Appendix Table 3). We used one category label (i.e., *Animal*) in the practice trial at the beginning of the experiment, and the remaining eight labels (i.e., *Bird*, *Color*, *Fish*, *Fruit*, *Insect*, *Instrument*, *Sport*, and *Weather*) in the interpolated-task phase of eight of the 16 critical trials. The number of letters in each exemplar and fragment varied such that in the practice phase, the exemplars included three to eight letters (M = 5.60, SD = 1.55) and the fragments included two to five letters (M = 3.20, SD= 1.01); in the actual experiment, the exemplars included three to ten letters (M = 5.84, SD = 1.83) and the fragments included one to seven letters (M = 3.62, SD = 1.30). All fragments included the first letter of the word, to allow for a high level of completion accuracy. We prerandomized the assignment of labels to blocks and kept the assignment constant across experimental formats. We chose exemplars from each category for fragment completion by ordering them randomly using a number generator and selecting the first 15 exemplars.

The materials for the two-back task were nine lists of individual lowercase letters from the English alphabet. We selected unique combinations of 15 letters for one practice list and for the eight lists that appeared in the critical trials of the experiment (for the complete material set, see Appendix Table 4). In each trial, we included five instances in which letters repeated after one nonrepeated letter (two-back target items), one to three instances in which the same letter appeared for consecutive items (one-back foils), and the remaining instances included letters that appeared once (foils). We prerandomized the assignment of lists to blocks and kept it constant across formats.

Procedure

We tested subjects individually in sessions scheduled for 2 h. Figure 1 displays a schematic of the procedure. Subjects first completed the dual-list free recall task and then completed the complex span tasks. We presented the tasks via computer using the E-Prime 2 software for the free recall task, and the E-Prime 1.2 software for the complex span tasks (Psychology Software Tools, Pittsburgh, PA, USA).

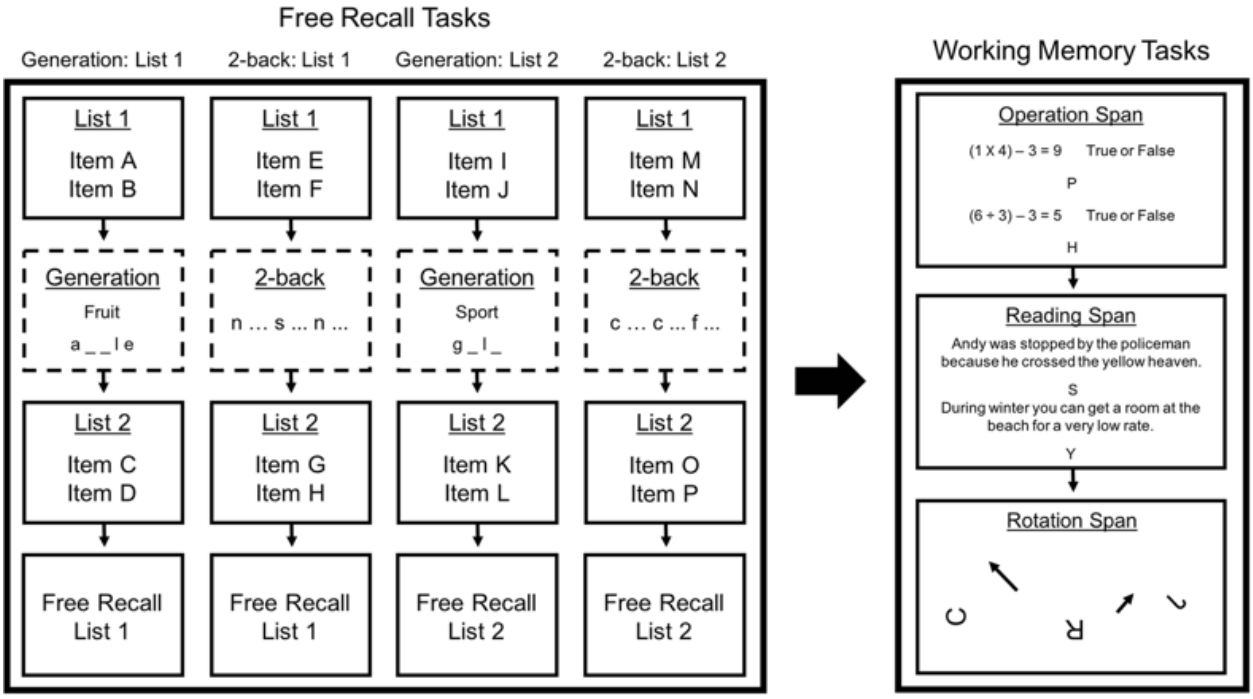

Figure 1. Schematic of the procedure for all tasks in the experiment. In the free recall phase (left), subjects completed four blocks of four trials (one block is displayed here) that each included one trial from each trial-type condition (16 total trials). Note that although the schematic above only includes two example items per list, the actual experiment included ten items per list. We abbreviated the schematic in this way to conserve space. Subjects then completed up to three complex span tasks (right)

Free recall. The dual-list free recall task began with a series of practice trials in which subjects separately practiced each task that they would eventually perform in the critical trials. During practice, subjects first completed a List 1 recall trial and then a List 2 recall trial. Both trials included two study lists with five items each, and there was no interpolated task between lists.

Subjects were encouraged to encode each list equally well as the instructions indicated that their task would be to recall one of the lists, but that they would not know which to recall from until after study. Subjects were instructed to read the words silently and to rate the pleasantness of each word on a scale from 1 (*low*) to 3 (*high*) by pressing the corresponding keys on a computer keyboard. Pleasantness ratings were included in order to improve attention during study and to keep study strategies as constant as possible.

Before each study list, a prompt displaying either "List 1" or "List 2" in yellow font appeared in the center of the screen against a black background for 2,750 ms, to indicate which list would be presented. A blank screen then appeared for 250 ms. Study items then appeared individually in the middle of screen in white font for 4,000 ms each. A response prompt that appeared below each study item read, "Pleasant" above the scale "low 1 2 3 high." Subjects rated each item by pressing the corresponding number key on the keyboard. When subjects responded, the prompt changed color from white to yellow to verify that the response had been registered. Subjects were instructed to continue studying each item until it disappeared. If subjects did not respond within 4,000 ms, then the program automatically advanced to the next item. A screen displaying only the pleasantness rating prompt appeared for 500 ms during the interstimulus intervals between items.

After the study phase, a retrieval cue appeared for 2,750 ms in yellow font in the middle of the screen, indicating the list from which subjects were to recall (i.e., "Recall from List 1," or "Recall from List 2"). The cue was followed by a blank screen for 250 ms. Subjects were instructed to recall words from the specified list in any order, to report any other response that came to mind, and to indicate whether each response was from the target list. Subjects typed each response onto the screen and pressed "Enter." Following each response, the prompt "1) correct, 2) incorrect" appeared, and subjects classified their response accuracy by pressing the corresponding number key. The recall period lasted 75 s. Subjects were instructed to continue reporting responses throughout the entire period and to wait silently when responses stopped coming to mind. Subjects were also instructed not to report the same item on subsequent recall attempts if it recurred in consciousness before another word came to mind.

After subjects completed the practice recall trials, they were given a practice two-back task. In this task, 15 individual letters appeared on the screen for 2,750 ms each, above a prompt that displayed "2-back?" above the options "Yes" and "No." Each letter was followed by a screen that displayed only the response prompt for 250 ms. Subjects were told that their task during this 3-s interval was to indicate whether the currently displayed letter was the same as the letter that had appeared two letters ago. Subjects responded "yes" or "no" aloud, and an experimenter recorded their responses.

After subjects had completed the practice two-back task, they were given a practice category exemplar generation task. In this task, a category label appeared in the middle of the screen above 15 individually presented word fragments of exemplars from that category. Each fragment appeared for 2,750 ms, followed by a screen that displayed only the category label for 250 ms. Subjects were told that during this 3-s interval, they should complete each fragment with an exemplar of the category. Subjects responded aloud, and an experimenter recorded their responses. Note that we equated both the study duration and response method in the generation

and two-back interpolated tasks, to better isolate the effects of retrieval type on free recall measures.

After practice, subjects began the critical trials of the free recall phase. The critical trials combined all the practiced tasks. The main differences between the critical and practice trials were that the critical-trial study lists each included ten items, and subjects completed one interpolated task between the lists per trial. The four within-subjects conditions were distributed evenly across four trial blocks (each block included one trial from each of the conditions). The presentation order of conditions within the trial blocks and items within the lists were both random.

After subjects had completed the last free recall trial, they were asked to report how frequently they had engaged in encoding strategies other than pleasantness ratings, on a Likert scale ranging from 0 (*never*) to 5 (*always*). When subjects provided a response above 0, another slide appeared asking them to enter any other strategies that they recalled using by typing them onto the screen. Subjects could provide as many responses as came to mind, and there was no time limit for responding. We did not have any specific hypotheses about how self-reported strategies would differ across subjects and whether they would create interactions with the independent variables of interest. We intended to use these data for exploratory analyses that could inform later experiments. However, after we finished data collection, we discovered that the responses were not recorded, due to a programming error.

Complex span tasks. Subjects completed three automated complex span tasks (Redick et al., 2012; Unsworth, Heitz, Schrock, & Engle, 2005), in the following order: (1) operation span (OPSAN), (2) reading span (RSPAN), and (3) rotation span (ROSPAN). In all three tasks, subjects immediately recalled short sequences of items presented on screen, with each memory item being preceded by an unrelated processing-task item that required a yes–no verification response (made via mouse click); following each sequence, all possible memory stimuli for that task appeared on screen (12 for OSPAN and RSPAN, 16 for ROSPAN), and subjects used a mouse to click the items from that trial in serial order. OSPAN required subjects to recall sequences of three to seven letters (presented for 1,000 ms each), interpolated with equations to verify; RSPAN required subjects to recall three to seven letters (presented for 1,000 ms each), interpolated with sentences to judge as sensible or nonsensical; and ROSPAN required subjects to recall two to five small versus large arrows radiating from center screen at one of eight angles (for 650 ms each), interpolated with rotated letters to judge as being normal or mirror-reversed.

All three tasks began with four practice trials of two or three memory items presented alone (with no processing task), then 15 practice trials of the processing task alone (with no memory items), and then three practice trials (of set size 2) of the memory and processing tasks combined. In the real task, the processing stimuli were presented until response, with a maximum duration equal to each subject's $M + 2.5$-SD processing time from the processing-only practice; if that maximum duration was reached on any trial, it was counted as a processing error, and the program advanced to the next item. Subjects completed three trials of each set size for each task.

Due to an error in the WMC programs, the accuracy of processing-task responses (e.g., math equations in OSPAN) was not properly recorded for some subjects. The purpose of this accuracy recording was to ensure that the subject paid proper attention to both tasks, as opposed to allocating all of their attention to the cues to be remembered. Despite the error, subjects should have believed that accuracy was being properly recorded, and they were told that the data from the session would not be used if their accuracy dropped below a threshold of 85%. Although we could not analyze the processing accuracy data or use them to identify potentially problematic subjects, the processing task retained its primary purpose of disrupting and limiting subjects' rehearsal of the memory items; importantly, any measurement error introduced by this inability to exclude noncompliant subjects should work against our finding associations of WMC with the other variables.

Results

In the following section, we test the model predictions by examining recall summary scores, recall conditionalized on input and output positions, and associations between recall measures and performance on the complex span tasks. The results are organized in the following way: First, we compare retrieval accuracy in the two interpolated tasks. Second, we examine free recall summary scores to determine how the interpolated-retrieval manipulation affected the accessibility and monitoring of correct recalls and intrusions at the list level. Third, we decompose correct recalls into serial-position (SP) curves and first-recall probability (FRP) functions across input positions, to determine how interpolated retrieval affected the accessibility of items in different list positions. Fourth, we compute standardized WMC composite scores for individual subjects and examine whether WMC was associated with the accessibility and monitoring of correct recall and intrusions. Finally, we verify that the associations between WMC and retrieval accuracy were comparable in the two interpolated tasks, and then examine SP curves and FRP functions conditionalized on WMC. We describe the rationale for each analysis below.

We set the significance level for all statistical tests at $\alpha = .05$. We modeled the effects of the experimental manipulations on free recall using linear mixed-effects models (LMMs) to account for subject variability (see Baayen, Davidson, & Bates, 2008; Jaeger, 2008) using the lmer function of the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) with R software (R Development Core Team, 2008). We then tested our hypotheses using the Anova function of the car package (Fox & Weisberg, 2011). Finally, we computed Bayes factors (*BF*) for experimental effects using functions from the BayesFactor package (Morey & Rouder, 2018) with R software.

Interpolated retrieval performance

Interpolated retrieval performance was high in both tasks, with both at approximately 90% accuracy, but performance was significantly higher in the two-back task ($M = .93$, $SD = .10$) than in the generation task ($M = .89$, $SD = .06$), $t(96) = 3.99$, $p < .001$, $d = 0.40$, $BF = 145.72$. These results suggest that the tasks were of numerically similar, but statistically different, difficulties.

Free recall summary scores

We coded free recall response types into four categories. *Correct recall* refers to responses recalled from the target list; *intratrial intrusions* refer to responses recalled from the nontarget list within the same trial as the target list; *prior-trial intrusions* refer to responses recalled from trials that preceded the current trial; and *extra-experiment intrusions* refer to responses that did not appear in the experiment. In the summary score analyses, we only analyze correct recalls and intratrial intrusions, because those responses should provide the most information about interpolated-retrieval effects on the accessibility and monitoring of temporally adjacent events. For each response type, we first examine differences in accessibility by comparing the overall number of responses produced in each condition. We then examine differences in monitoring accuracy by comparing the numbers of responses classified as correct.

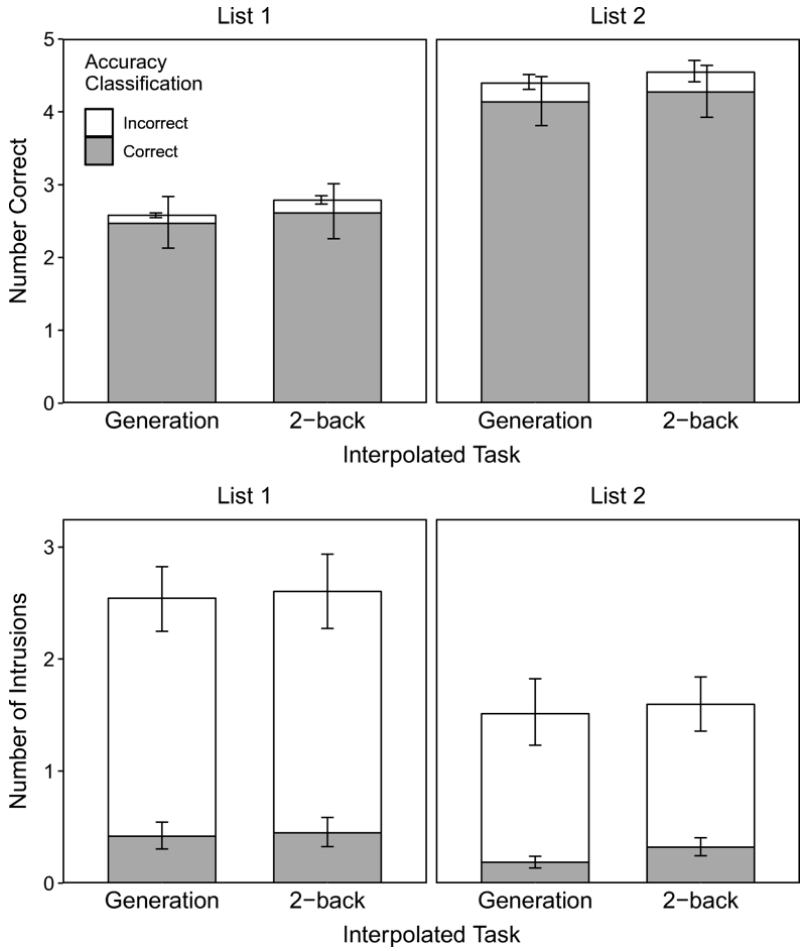

Figure 2. Mean proportions of correctly recalled responses (top panels) and intratrial intrusions (bottom panels) as a function of trial type, interpolated task, and accuracy classification. The total possible number of responses for each measure was 10. Overall response production is displayed as the total height of each bar (including both the white and gray bars), and the frequency of responses being classified as "correct" is displayed as the gray bars. Error bars showing bootstrap 95% confidence intervals are displayed for each type of accuracy classification

Correct recall. To determine whether interpolated retrieval from LTM decreased List 1 accessibility and increased List 2 accessibility (relative to STM retrieval), as had been observed

in earlier studies (e.g., Divis & Benjamin, 2014; Jang & Huber, 2008), we compared the numbers of correct responses produced per trial across within-subjects conditions (Fig. 2, top panels [total height of bars]). We fitted to these data an LMM and a Bayes factor analysis of variance ($ANOVA^{BF}$), both with trial type and interpolated task as fixed effects and subjects as a random effect. The LMM indicated a significant effect of trial type, $\chi^2(1) = 342.55$, $p < .001$ ($BF = 7.80 \times 10^{46}$), no significant effect of interpolated task, $\chi^2(1) = 2.56$, $p = .11$ ($BF = 0.19$), and no significant Trial Type × Interpolated Task interaction, $\chi^2(1) = 0.30$, $p = .59$ ($BF = 0.22$). These results show that List 2 items were more accessible than List 1 items, presumably because the test context was more similar to the List 2 than to the List 1 study context. However, the interpolated-task manipulation did not appear to affect response accessibility in either study list.

We further examined whether the predicted effects of interpolated retrieval would arise when considering only correct responses that subjects classified as such. This measure of correct recall is more similar to that used in standard free recall, which has revealed interpolated-task effects on response accessibility (e.g., Divis & Benjamin, 2014; Jang & Huber, 2008). We examined whether the interpolated-task manipulation affected correct recall monitoring by comparing the numbers of responses classified as "correct" across within-subjects conditions (Fig. 2, top panel [gray bars]) using the same model types as in the previous analyses. The LMM indicated a significant effect of trial type, $\chi^2(1) = 270.12$, $p < .001$ ($BF = 3.53 \times 10^{39}$), no significant effect of interpolated task, $\chi^2(1) = 1.63$, $p = .20$ ($BF = 0.17$), and no significant Trial Type × Interpolated Task interaction, $\chi^2(1) = 0.03$, $p = .87$ ($BF = 0.15$). These results replicated the pattern of response accessibility, indicating that the manipulation of interpolated retrieval did not affect correct recall monitoring.

Intratrial intrusions. We further examined the effects of interpolated retrieval by comparing intratrial intrusion accessibility across trial-type conditions (Fig. 2, bottom panel [total height of bars]). We fitted to these data an LMM and an $ANOVA^{BF}$, with trial type and interpolated task as fixed effects and subjects as a random effect. The LMM indicated a significant effect of trial type, $\chi^2(1) = 133.77$, $p < .001$ ($BF = 2.49 \times 10^{22}$), no significant effect of interpolated task, $\chi^2(1) = 0.56$, $p = .45$ ($BF = 0.14$), and no significant Trial Type × Interpolated Task interaction, $\chi^2(1) = 0.01$, $p = .94$ ($BF = 0.16$). These results showed that more intratrial intrusions from List 2 onto List 1 were produced than were intratrial intrusions from List 1 onto List 2. This was also presumably due to the test context being more similar to the List 2 than to the List 1 context. Consistent with the results from correct recall, we found no interpolated-retrieval effects on intrusion accessibility.

As with the analyses of correct recall, we also examined interpolated-retrieval effects on intratrial intrusion monitoring by comparing the numbers of intratrial intrusions classified as "correct" (Fig. 2, bottom panel [gray bars]) across within-subjects conditions using the same models as in the previous analyses. The LMM indicated a significant effect of trial type, $\chi^2(1) = 16.37$, $p < .001$ ($BF = 248.55$), no significant effect of interpolated task, $\chi^2(1) = 3.40$, $p = .07$ ($BF = 0.53$), and no significant Trial Type × Interpolated Task interaction, $\chi^2(1) = 1.30$, $p = .25$ ($BF = 0.28$). These results show that subjects classified as "correct" more intrusions from List 2 onto List 1 recall than the reverse, which was likely due to the greater similarity between the test and List 2 contexts. However, there were no monitoring differences between the interpolated-

task conditions. Collectively, the results from these summary-score measures are inconsistent with the prediction that interpolated LTM retrieval should accelerate internal context change.

Free recall dynamics

Serial-position curves. Although the summary scores above can be used to determine whether the interpolated-task manipulation influenced response accessibility, and thus internal context change, this approach obscures potential differences across items. There might not have been accessibility differences between interpolated-task conditions in the analyses above because those differences emerged only for a subset of the items. In particular, accessibility differences might only have been present in early List 1 and late List 2 positions, because those items were the first and last to appear in the contextual stream. We therefore conditionalized total correct recall production on input position (i.e., we created SP curves) and compared this measure of accessibility between the interpolated-task conditions. Note that the following analyses were exploratory, since we did not originally anticipate selective accessibility differences. To analyze the SP curves, we fitted LMMs and ANOVA^{BF}s separately to the data in each trial type condition, with interpolated task and input position as fixed effects and subjects as a random effect.

Figure 3 (left panel) displays the SP curves for the List 1 condition. The LMM indicated a significant effect of interpolated task, $\chi^2(1) = 4.32$, $p = .04$ ($BF = 0.42$), a significant effect of input position, $\chi^2(9) = 92.30$, $p < .001$ ($BF = 1.61 \times 10^{12}$), and no significant Interpolated Task × Input Position interaction, $\chi^2(1) = 7.47$, $p = .59$ ($BF < .01$). These results indicated shallow primacy gradients that were similar in each interpolated-task condition. However, there was a suggestive but nonsignificant trend toward a steeper primacy gradient in the early positions of List 1 for the two-back than for the generation condition. To foreshadow, this interaction might only be detectable when considering individual differences in context processing (measured here as WMC).

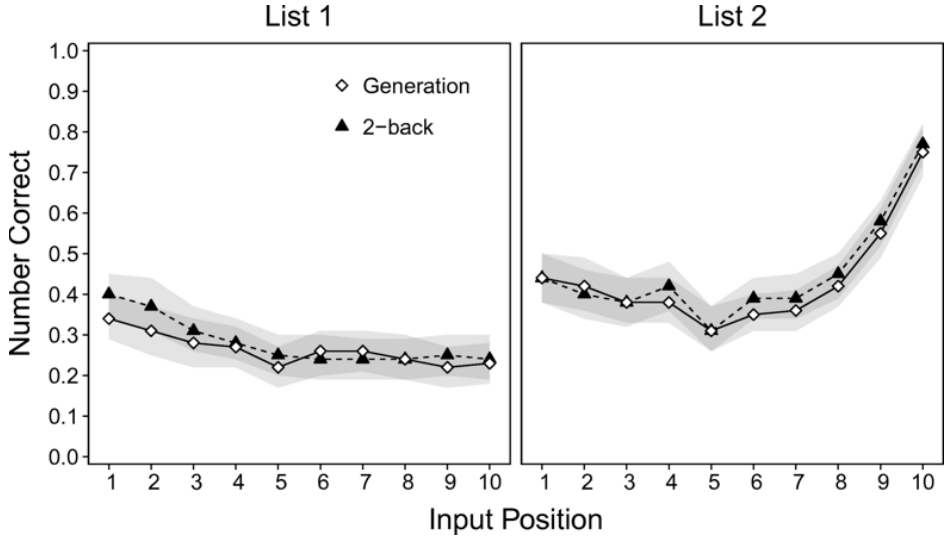

Figure 3. Serial-position curves, displaying the mean numbers of correct responses output per trial as a function of trial type, interpolated task, and input position. Shaded regions represent bootstrap 95% confidence intervals

Figure 3 (right panel) displays the SP curves in the List 2 condition. The LMM indicated no significant effect of interpolated task, $\chi^2(1) = 1.03$, $p = .31$ ($BF = 0.08$), a significant effect of input position, $\chi^2(9) = 483.71$, $p < .001$ ($BF = 3.11 \times 10^{83}$), and no significant Interpolated Task × Input Position interaction, $\chi^2(9) = 3.46$, $p = .94$ ($BF < .01$). These results showed shallow primacy and steep recency effects that were nearly identical in both interpolated-task conditions. These patterns suggest that the interpolated-task manipulation did not affect correct recall accessibility in any of the List 2 input positions.

First-recall probabilities. Following the analyses of SP curves, accessibility differences might also be indicated by examining the list positions from which subjects initiated their first retrieval attempt of the recall period. For example, according to the buffer model (Lehman & Malmberg, 2009, 2013), the magnitude of the FRPs for the first item studied indicates the extent to which context reinstatement was at least partially used to initiate retrieval. FRP functions may therefore be a sensitive measure of interpolated-retrieval effects on context change, because differences in FRPs for initially studied items should be obtained when an interpolated-retrieval manipulation affects the similarity between the study and test contexts. As with the SP curves, we performed exploratory analyses on FRP functions using the same model types as for the SP curves.

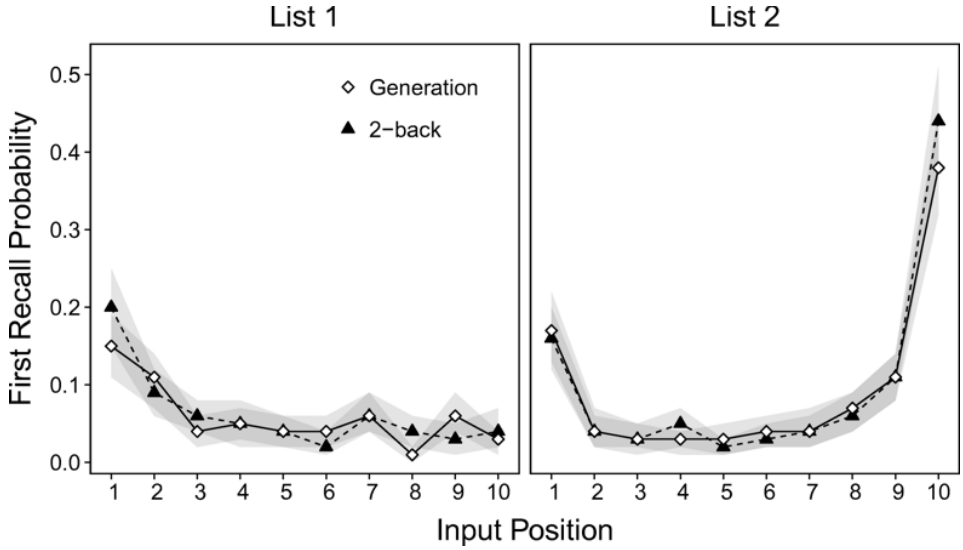

Figure 4. First-recall probabilities, displaying the mean numbers of correct responses in the first output position per trial as a function of trial type, interpolated task, and input position. Shaded regions represent bootstrap 95% confidence intervals

Figure 4 (left panel) displays FRP functions in the List 1 condition. The LMM indicated no significant effect of interpolated task, $\chi^2(1) = 0.35$, $p = .55$ ($BF = 0.06$), a significant effect of input position, $\chi^2(9) = 211.46$, $p < .001$, ($BF = 1.78 \times 10^{35}$), and no significant Interpolated Task × Input Position interaction, $\chi^2(9) = 13.87$, $p = .13$, ($BF = .02$). Subjects initiated List 1 recall most often from the primacy positions, which generally replicates response initiation patterns showing primacy but not recency effects in delayed free recall tests (e.g., Kahana, Howard, Zaromb, & Wingfield, 2002). Although subjects may have used context reinstatement to initiate retrieval, the extent to which they could access items in the first position did not differ between

interpolated tasks. It remained possible, however, that this pattern could differ depending on WMC; we explore that possibility below.

Given that the SP curves in the List 2 condition were nearly identical in each interpolated-task condition, we did not expect to see differences in the List 2 FRP functions. However, we wanted to verify that the patterns of response initiation in that condition aligned with earlier studies using immediate free recall. Figure 4 (right panel) displays the FRP functions in the List 2 condition. The LMM indicated no significant effect of interpolated task, $\chi^2(1) = 0.27$, $p = .60$ ($BF = 0.05$), a significant effect of input position, $\chi^2(9) = 1{,}050.55$, $p < .001$ ($BF = 2.72 \times 10^{170}$), and no significant Interpolated Task × Input Position interaction, $\chi^2(9) = 8.66$, $p = .47$ ($BF < .01$). Primacy effects were similarly smaller than recency effects across the interpolated-task conditions. These functions replicated earlier findings from immediate free recall tests showing smaller primacy than recency effects (e.g., Lehman & Malmberg, 2013). Subjects thus alternated their response initiation between the first and last positions of List 2 across trials, but they initiated retrieval from the last positions more often. According to both the buffer and WMC models, such response initiation patterns could differ on the basis of WMC. We examine this possibility below.

Individual differences in WMC

We assessed WMC for each subject by computing the sum of the items correctly recalled in their original position for each complex span task. The theoretical maximum scores were 75 for OSPAN and RSPAN, and 42 for ROSPAN. Of the 95 subjects included in these analyses, 90 had completed all measures, three had completed the OSPAN and RSPAN measures, and two had completed the RSPAN and ROSPAN measures. The scores for each of the complex spans were: OSPAN ($M = 50.81$, $SD = 15.20$, range = 7–75, $N = 93$), RSPAN ($M = 45.54$, $SD = 15.45$, range = 4–72, $N = 95$), and ROSPAN ($M = 22.90$, $SD = 9.17$, range = 0–40, $N = 92$). The correlations among span scores were: OSPAN × RSPAN, $r(91) = .71$, $CI = [.59, .80]$, $p < .001$; OSPAN × ROSPAN, $r(88) = .45$, $CI = [.26, .60]$, $p < .001$; RSPAN × ROSPAN, $r(90) = .45$, $CI = [.28, .60]$, $p < .001$.

Table 1. Unstandardized and standardized working memory capacity (WMC) span scores as a function of task and WMC tercile

WMC (Tercile)	Unstandardized (Raw) Scores			Standardized (*Z*) Scores		
	OSPAN	**RSPAN**	**ROSPAN**	**OSPAN**	**RSPAN**	**ROSPAN**
High (3rd)	62.31 (6.42)	58.06 (8.28)	30.19 (5.52)	0.76 (0.42)	0.81 (0.54)	0.79 (0.60)
Medium (2nd)	52.87 (8.37)	48.68 (8.26)	23.69 (7.06)	0.14 (0.55)	0.20 (0.54)	0.09 (0.77)
Low (1st)	36.40 (15.85)	29.97 (12.92)	14.65 (7.04)	– 0.95 (1.04)	– 1.01 (0.84)	– 0.90 (0.77)

Mean scores are displayed in each cell, and standard deviations are displayed in parentheses. OSPAN = operation span; RSPAN = reading span; ROSPAN = rotation span

We created composite WMC scores for each subject by standardizing and averaging the scores from each task. Composite scores for subjects who did not complete all WMC tasks were the average of the tasks they completed. We used the composite scores in a series of correlational analyses to test theoretical functions of WMC. We also divided subjects into three terciles based on the WMC composite scores. This allowed us to examine differences in the SP curves and FRP functions based on WMC. The upper (3rd) and lower (1st) terciles each included 32 subjects, and

the middle tercile (2nd) included 31 subjects. Table 1 displays descriptive statistics for the unstandardized and standardized scores across terciles.

WMC, response production, and monitoring

As we described above, two theoretical functions of WMC are to reinstate context (Unsworth & Engle, 2007) and to monitor the source of a retrieved context (Unsworth & Brewer, 2010). We assessed evidence for these functions by conducting a series of between-subjects correlations examining the relations between WMC and the production and monitoring of free recall responses. We measured WMC as standardized WMC composite scores, response production as the mean number of responses produced per trial, and monitoring accuracy as the proportion of correctly classified responses per trial. If WMC serves to reinstate context, then it should correlate positively with the production of correct recalls and negatively with the production of intrusions from other sources (e.g., Unsworth, 2016; Unsworth & Brewer, 2010). In addition, if WMC facilitates monitoring of the source of a retrieved context, it should correlate positively with the proportion of correctly classified responses (e.g., Unsworth & Brewer, 2010). We tested these predictions by computing bivariate Pearson correlations between WMC and monitoring accuracy on four recall measures (correct recall, intratrial intrusions, prior-trial intrusions, and extra-experimental intrusions). We assessed the reliability of each recall measure by computing Cronbach's alpha using each of the four trial blocks as "items" in the analysis (see Table 2).

Table 2. Cronbach's alpha reliability estimates for free recall measures included in the bivariate correlations

Recall measure	Trial type	Interpolated	Response type			
			Correct	ITI	PTI	EEI
Production	List 1	Generation	.74 [.66, .83]	.69 [.59, .79]	.65 [.54, .76]	.82 [.76, .88]
		Two-back	.76 [.69, .84]	.80 [.73, .87]	.73 [.65, .80]	.79 [.72, .86]
	List 2	Generation	.74 [.65, .82]	.73 [.64, .82]	.83 [.77, .88]	.85 [.80, .90]
		Two-back	.73 [.64, .82]	.71 [.61, .80]	.77 [.71, .83]	.69 [.58, .79]
Monitoring	List 1	Generation	– .13 [– .50, .23]	.74 [.66, .83]	.36 [.12, .61]	.71 [.59, .82]
		Two-back	.21 [– .05, .47]	.54 [.38, .70]	.64 [.50, .78]	.68 [.55, .81]
	List 2	Generation	.72 [.63, .81]	.75 [.67, .84]	.70 [.57, .82]	.80 [.72, .89]
		Two-back	.85 [.80, .90]	.49 [.31, .66]	.83 [.76, .91]	.70 [.57, .82]

95% confidence intervals are displayed in brackets. ITI = intratrial intrusions; PTI = prior-trial intrusions; EEI = extra-experiment intrusions

WMC and response production. Figure 5 displays scatter plots showing the associations between WMC and response production for each measure in each within-subjects condition. The reliability estimates for the following recall measures were all adequate (Table 2, top panel). Figure 5 (top left panels) shows that WMC was modestly (positively) associated with correct recall production, which is consistent with the assumption that one function of WMC is to reinstate the target-list context (Unsworth & Engle, 2007). However, the modest associations also suggest that much of the individual variation is independent of WMC. The remaining plots indicated little, if any, association between WMC and intrusion production, except for a modest positive association for intratrial intrusions in the List 2/generation condition. The general pattern of association between WMC and intrusions is inconsistent with the prediction of a negative association between WMC and intrusion production (cf. Unsworth & Brewer, 2010).

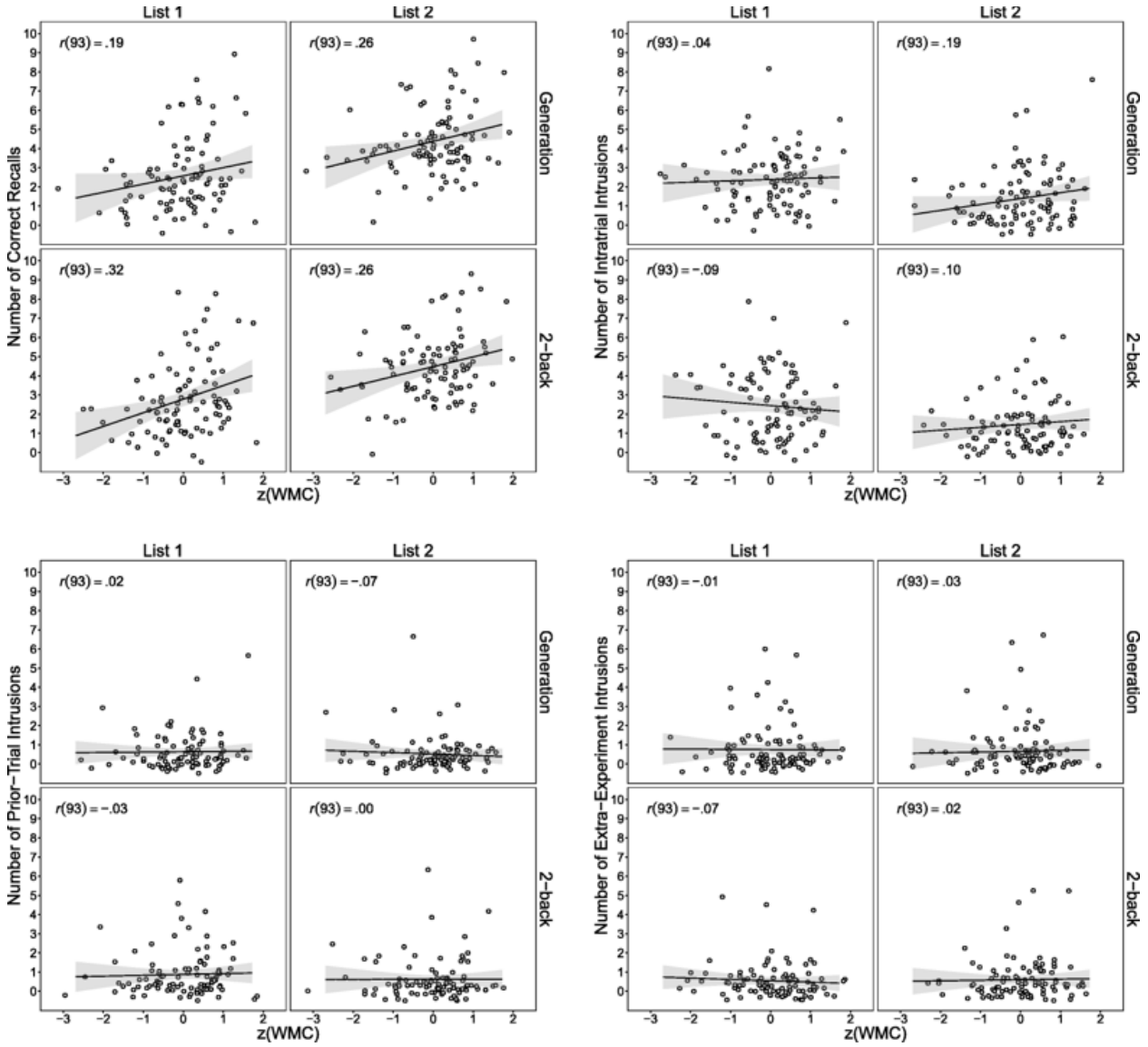

Figure 5. Scatter plots showing the relationship between individual differences in total response production (for correct recall and each intrusion type) and standardized working memory capacity (WMC) composite scores as a function of trial type and interpolated task. Shaded regions represent 95% confidence intervals

WMC and response monitoring. Figure 6 displays scatter plots showing the association between WMC and monitoring for each measure in each within-subjects condition. In contrast to previous analyses, reliability estimates varied in their adequacy across conditions (Table 2, bottom panel). We note specific instances of inadequate reliability below. Figure 6 (top left panels) shows that, in general, WMC was modestly (positively) associated with correct recall monitoring. However, these associations should be interpreted cautiously, because there were ceiling effects for most subjects, and reliability was inadequate in the List 1 conditions. The remaining panels show that WMC was also, in general, modestly (positively) associated with intrusion monitoring for most conditions, except that some conditions showed small positive associations, and one condition showed a small negative association. As with the correlations

involving correct recall monitoring, the associations between intrusion monitoring and WMC should be interpreted cautiously, given that the distribution of monitoring scores appeared to deviate from normalcy (i.e., many subjects were at the ceiling and floor). In addition, reliability was inadequate for intratrial intrusions in the List 1 and List 2/two-back conditions, and for prior-trial intrusions in the List 1/generation condition. Despite these caveats, several of these patterns are consistent with the assumption that one function of WMC is to monitor the source of a retrieved context (Unsworth & Brewer, 2010); however, as with the production measure, much of the individual variation is independent of WMC.

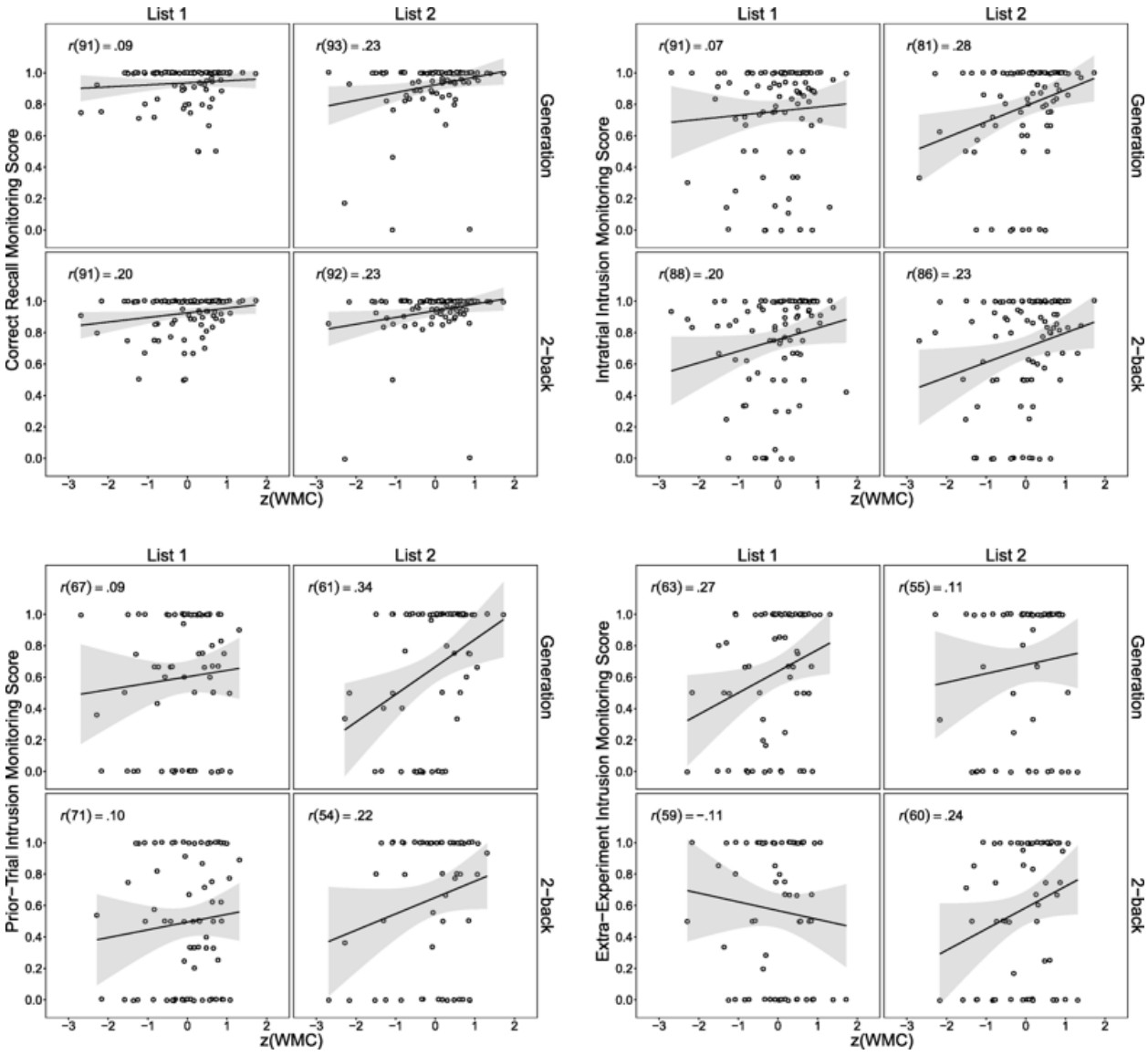

Figure 6. Scatter plots showing the relationship between individual differences in the monitoring of intrusions and standardized working memory capacity (WMC) composite scores as a function of response type, trial type, and interpolated task. Shaded regions represent 95% confidence intervals

WMC and free recall dynamics

The associations between WMC and measures of response production and monitoring in prior analyses were somewhat consistent with the assumption that WMC is involved in reinstating and monitoring context. However, the clearest associations were those showing modest positive relationships between WMC and correct recall production. Any interactive effects between WMC and the interpolated-retrieval manipulation should therefore be most apparent in the measure of correct recall production. In the following analyses, then, we decomposed correct recall production into SP curves and FRP functions and examined whether the shapes of those functions differed across the WMC terciles.

WMC and interpolated-retrieval performance. Before examining how WMC was related to the SP curves and FRP functions, we first determined whether retrieval accuracy in each interpolated task was comparably sensitive to individual differences in WMC. We did this by computing bivariate Pearson correlations between WMC and retrieval accuracy in each task (category exemplar generation and two-back task). The outcome of this analysis is important for interpreting any potential interactive effects of WMC and interpolated retrieval. Specifically, accessibility differences between interpolated-task conditions could be interpreted as evidence for individual differences in the rates of context change between lists. Alternatively, such differences could be attributed to individual differences in the ability to rehearse List 1 items while completing the interpolated task. The latter account would then suggest that context change did not play a primary role in accessibility differences. Evidence against a rehearsal-based account would be shown by interpolated-task performance varying across WMC in a similar manner in both tasks, since this would indicate that WMC did not confer differential rehearsal benefits in one task (cf. Gardiner, Thompson, & Maskarinec, 1974).

This consideration has theoretical implications, because one could predict that higher WMC would enable task completion more for the two-back task than for category exemplar generation. This has implications for interpreting differences in the free recall functions across WMC terciles, should they occur. Specifically, if WMC does *not* selectively advantage two-back task completion and the shape of the List 1 recall functions were to indicate greater accessibility of early List 1 items in the two-back condition for higher-WMC subjects, this would suggest that the context change account of differences in List 1 accessibility resulting from interpolated LTM retrieval was more plausible than a rehearsal-based account.

Before computing the correlations between WMC and interpolated-task performance, we assessed the reliability of the measurements of interpolated-retrieval accuracy by computing Cronbach's alpha for each interpolated task. We collapsed across the trial-type condition and treated the four trial blocks as items. Each measure was adequately reliable: generation (α = .81, *CI* = [.76, .87]); two-back (α = .94, *CI* = [.92, .96]). The scatter plots showing the association between mean interpolated-task performance and WMC (Fig. 7) showed a numerically larger WMC correlation for the generation than for the two-back task. However, given that both correlations were positive, with interpolated-task performance being better for higher- than for lower-WMC subjects, any differences in List 1 response accessibility across WMC groups was unlikely to reflect differential List 1 rehearsal across interpolated tasks. Moreover, the numerically weaker WMC correlation in the two-back condition appears to be driven by one

subject with an especially low score; with this subject removed, the correlation increased to r = .35, which is closer still to the magnitude of the WMC–generation correlation (r = .40).

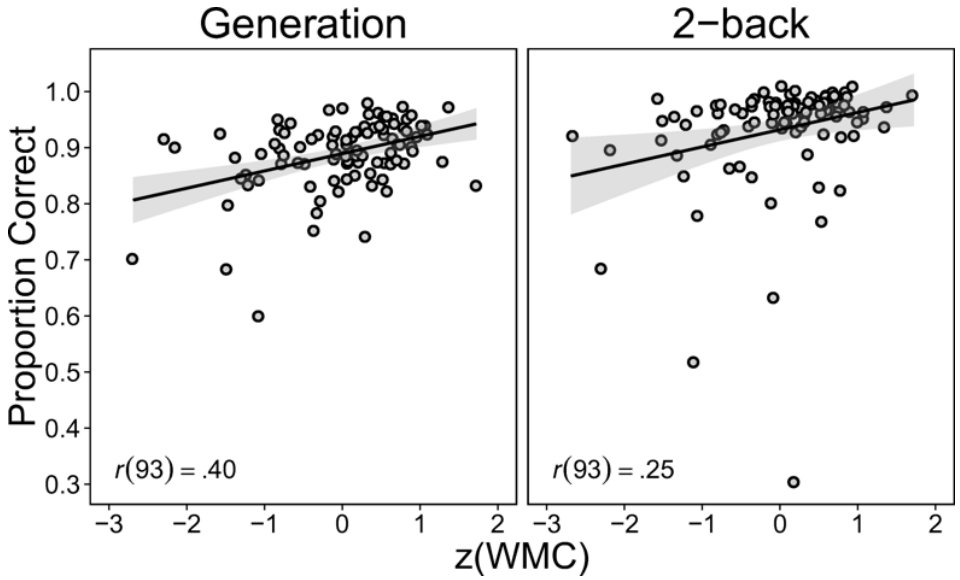

Figure 7. Scatter plots showing the relationship between individual differences in interpolated-task performance and standardized working memory capacity (WMC) composite scores as a function of interpolated task. Shaded regions represent 95% confidence intervals

WMC and serial-position curves. Examination of the SP curves aggregated across all subjects in the List 1 condition (Fig. 3) suggested that response accessibility in early List 1 positions may have been higher in the two-back than in the generation condition. In the following analyses, we examined whether these differences were amplified for higher-WMC subjects, which could reflect enhanced context utilization. We also examined whether WMC was associated with the SP curves in the List 2 condition. On the basis of the finding above that the overall SP curves in List 2 were nearly identical in the two interpolated-task conditions, we did not expect that WMC would interact with the interpolated-task manipulation. However, we expected differences in the List 2 SP curves, based on the finding that higher-WMC subjects show larger primacy effects in immediate free recall than do lower-WMC subjects (e.g., Unsworth & Engle, 2007). Note that the analyses of SP curves were largely exploratory.

Did higher-WMC subjects show more pronounced differences in List 1 primacy effects in the two-back versus the generation conditions than did lower-WMC subjects? The buffer model predicts overall primacy effects in List 1 recall, because fewer items occupy the buffer during the first portion of the list, which allows stronger context-to-item associations to be established. Following this reasoning, higher-WMC subjects should show larger and more prolonged primacy effects, because they can hold more items in the buffer than can lower-WMC subjects. Furthermore, if one assumes that context change was accelerated by interpolated LTM retrieval, this would lead higher-WMC subjects to show a larger primacy advantage in the two-back than in the generation condition than would lower-WMC subjects. This would result from the category exemplar generation task creating greater contextual change, leading to reduced accessibility of the early List 1 items following interpolated LTM retrieval. WMC theory also predicts these effects, by assuming that higher-WMC subjects utilize context more effectively

than lower-WMC subjects, leading to more rapid context updating from interpolated LTM retrieval.

We analyzed the SP curves in each trial type condition using LMMs and ANOVA^{BF}s with fixed effects of WMC, input position, and interpolated task, as well as a random effect of subjects. To limit redundancy, we only report results involving interactions with the fixed effect of WMC. Figure 8 (top panels) displays the SP curves in the List 1 condition in each of the three WMC terciles. The LMM indicated no significant WMC × Input Position interaction, $\chi^2(18) = 24.72$, $p = .13$ ($BF < 0.01$), a significant WMC × Interpolated Task interaction, $\chi^2(2) = 6.64$, $p = .04$ ($BF = 0.26$), and a significant WMC × Input Position × Interpolated Task interaction, $\chi^2(18) = 38.60$, $p = .003$ ($BF = 2.69$). Consistent with predictions from the buffer and WMC models, these results indicated larger primacy effects in the two-back than in the generation condition for high-WMC subjects, with this effect being smaller for middle-WMC subjects and absent for low-WMC subjects. Note, however, that the *BF* value for the three-way interaction suggested only weak evidence for the alternative hypothesis, and so should be interpreted with caution.

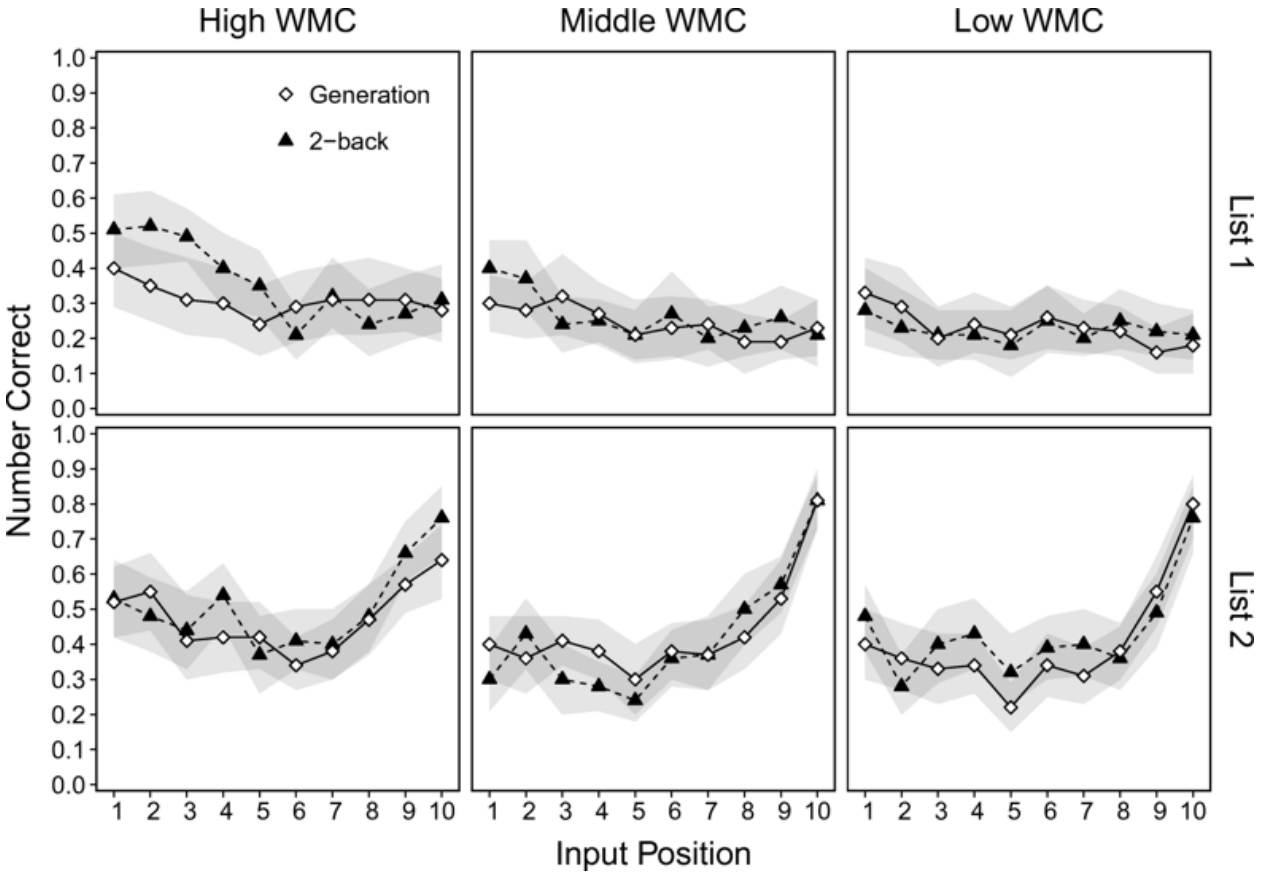

Figure 8. Serial-position curves as a function of working memory capacity (WMC) tercile, trial type, and interpolated task. Shaded regions represent bootstrap 95% confidence intervals

Given that the SP curves in the List 2 condition aggregated across all subjects appeared identical for each interpolated-task condition (see Fig. 3), we did not expect the interpolated-task manipulation to interact with WMC. However, we examined the shapes of the SP curves across

WMC groups to confirm the prediction from WMC theory (Unsworth & Engle, 2007), that higher-WMC subjects should show larger primacy effects than lower-WMC subjects.

Figure 8 (bottom panels) displays the SP curves in the List 2 condition in each of the three WMC terciles. The LMM indicated a significant WMC × Input Position interaction, $\chi^2(18) = 51.43$, $p < .001$ ($BF = 47.04$), no significant WMC × Interpolated Task interaction, $\chi^2(2) = 4.26$, $p = .12$ ($BF = 0.06$), and a significant WMC × Interpolated Task × Input Position interaction, $\chi^2(18) = 30.02$, $p = .04$ ($BF = 0.19$). The WMC × Input Position interaction indicates that the recency gradient was shallowest for high-WMC subjects, which is consistent with results showing that these subjects initiated retrieval from earlier positions than did the lower-WMC subjects (Unsworth & Engle, 2007). The three-way interaction suggests differences in recall probabilities between the interpolated-task conditions in the recency portion for high-WMC subjects, and in the early middle portions for the other groups, but the low *BF* value suggests a cautious interpretation.

WMC and first-recall probability functions. As with the analyses of individual differences in SP curves above, the analyses of individual differences in FRP functions were largely exploratory. However, the buffer and WMC models make testable predictions about variations in these functions based on WMC. According to the buffer model, context-to-item associations should be strongest for the first few items in a list, and subjects with larger capacity buffers should establish more effective context-to-item associations. If higher-WMC subjects form stronger context-to-item associations, they should show larger FRP primacy effects in the List 1 condition than should lower-WMC subjects. The prediction from the buffer model could also be extended to predict larger FRP primacy effects in the two-back than in the generation condition for higher- than for lower-WMC subjects. This is because the accelerated context change that is assumed to occur in the generation condition, and to be further amplified for higher-WMC subjects, would exaggerate the difference between the study and test contexts, thus limiting context reinstatement. Although WMC theory does not make direct predictions about differences in FRP primacy on delayed-recall tests, it does predict a broader distribution of response initiation for higher- than for lower-WMC subjects in immediate free recall. This could be extended to predict differences in FRP primacy based on WMC, but the direction of these differences is unclear.

As with the SP curves, we analyzed the FRP curves in each trial-type condition using LMMs and ANOVA^{BF}s with fixed effects of WMC, input position, and interpolated task, as well as a random effect of subjects. To limit redundancy, we only report results involving interactions with the fixed effect of WMC. Figure 9 (top panels) displays the FRP functions for the List 1 condition across the three WMC terciles. We examined the FRP curves in the List 1 condition differently than we examined the SP curves, due to our interest in early list positions. Here we focused on the first three input positions, because interpolated-task differences in the List 1 SP curves were most pronounced for the high-WMC subjects there, which suggested that WMC modulated context utilization. The LMM indicated no significant WMC × Input Position interaction, $\chi^2(4) = 3.81$, $p = .43$ ($BF = 0.14$), a significant WMC × Interpolated Task interaction, $\chi^2(2) = 8.73$, $p = .01$ ($BF = 1.07$), and a significant WMC × Input Position × Interpolated Task interaction, $\chi^2(4) = 15.77$, $p = .003$ ($BF = 16.34$). High-WMC subjects initiated recall from the beginning of List 1 more often in the two-back than in the generation

condition, and this trend was reversed for low-WMC subjects. These results are consistent with the interpretation that high-WMC subjects utilized context more effectively in the two-back than in the generation condition.

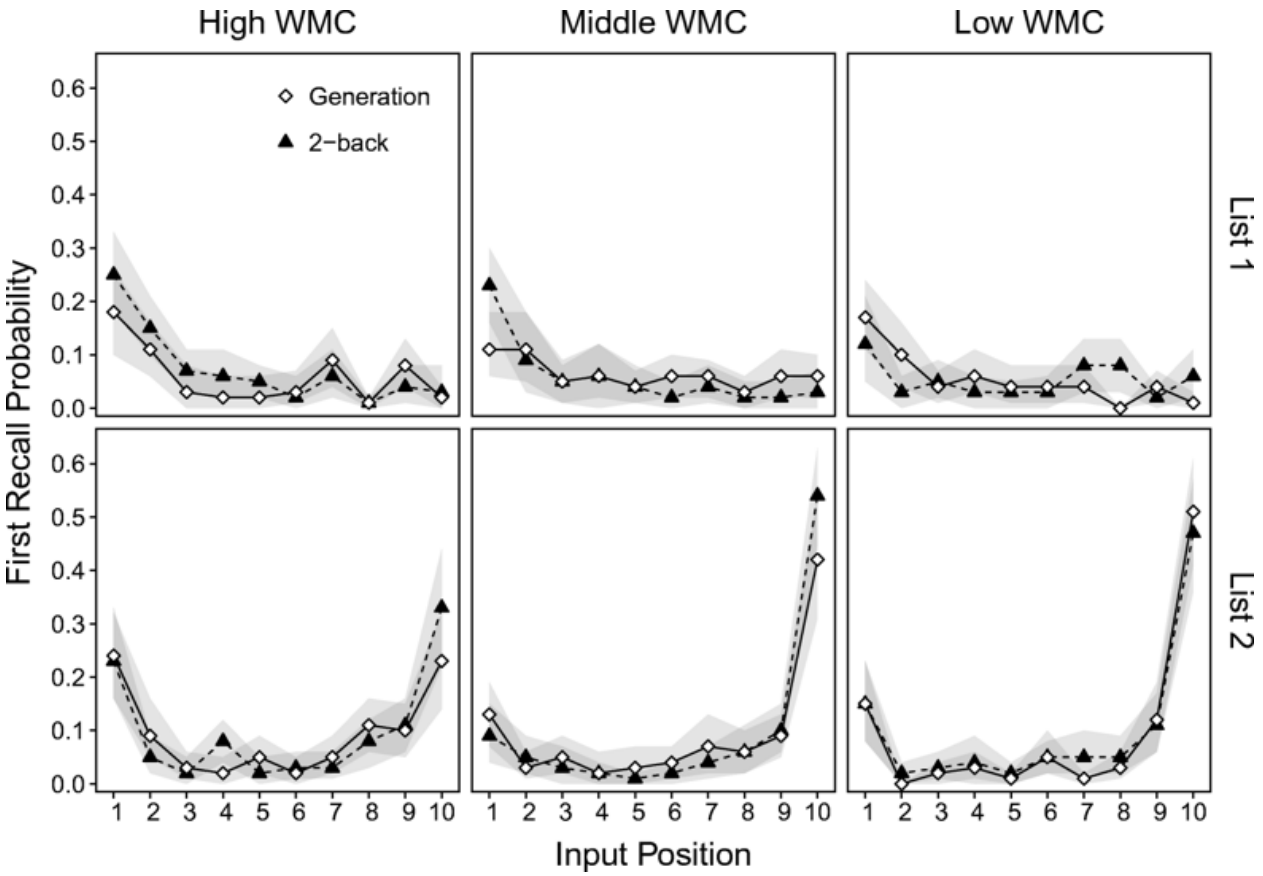

Figure 9. First-recall functions as a function of working memory capacity (WMC) tercile, trial type, and interpolated task. Shaded regions represent bootstrap 95% confidence intervals

For the List 2 condition, the buffer and WMC models diverge in their predictions about retrieval initiation patterns in immediate free recall. The buffer model predicts that subjects with lower capacities for rehearsal (i.e., lower WMC) should initiate retrieval from positions at the beginning of the list more than higher-WMC subjects do, because for lower-WMC subjects, relatively few end-of-list items would be available. In contrast, WMC theory predicts that higher-WMC subjects should distribute their recall initiation across more list items than would lower-WMC subjects, because higher-WMC subjects maintain more information in working memory.

Figure 9 (bottom panels) displays the FRP functions for the List 2 condition across the three WMC groups. We fitted an LMM to all input positions in order to examine differences in the distributions of initial responses across the entire list. The model indicated a significant WMC × Input Position interaction, $\chi^2(18) = 122.57$, $p < .001$ ($BF = 6.87 \times 10^{13}$), no significant WMC × Interpolated Task interaction, $\chi^2(2) = 0.05$, $p = .97$ ($BF = 0.01$), and no significant WMC × Input Position × Interpolated Task interaction, $\chi^2(18) = 19.35$, $p = .37$ ($BF = 0.01$). These results are more consistent with WMC theory than with the buffer model, in showing that higher-WMC

subjects distributed their response initiation across more input positions. We consider the theoretical implications of this collection of results below.

Discussion

In the present experiment we examined the effects of interpolated retrieval from LTM and STM on list isolation in dual-list free recall, and assessed whether these effects interacted with individual differences in WMC. We used an EFR procedure to evaluate any effects of interpolated retrieval and WMC on response accessibility and monitoring. Interpolated LTM retrieval did not lead to differences in correct recall or in intratrial intrusions in summary score measures of accessibility or monitoring. From a context change perspective, this suggested that interpolated LTM retrieval did not accelerate context change relative to interpolated STM retrieval. WMC was moderately associated with correct recall production and, in some cases, intrusion monitoring. These findings are consistent with the suggestion that WMC serves to reinstate and monitor context, but much of the individual variation in WMC was independent of these measures. Exploratory analyses of the List 1 condition revealed greater primacy effects in SP curves and FRP functions when higher-WMC subjects received interpolated STM retrieval. In contrast to summary scores, these effects on early List 1 items suggested that higher-WMC subjects experienced greater context change following interpolated LTM than interpolated STM retrieval. Finally, exploratory analyses of the List 2 condition showed that higher-WMC subjects recalled from (SP curves) and initiated retrieval from (FPR functions) earlier input positions than did lower-WMC subjects. This is consistent with the model assumption that higher-WMC subjects can maintain more information in temporary storage (WMC theory). We discuss these findings in more detail and consider their theoretical implications below.

Retrieval-induced list isolation

Interpolated task manipulations that decrease the accessibility of previous lists and increase the accessibility of later lists may have such an influence on list isolation by inducing internal context change. This perspective aligns with the assumption of retrieved context models that contextual drift across items is accelerated by retrieval events (e.g., Howard & Kahana, 2002; Lohnas et al., 2015; Polyn et al., 2009). We assumed, on the basis of earlier suggestions, that the primary difference between the interpolated tasks in the present experiment were that the category exemplar generation task relied more on LTM retrieval than did the two-back task, and that both relied on STM (e.g., Jang & Huber, 2008). We also thought it unlikely that our interpolated tasks differed in their suppression of List 1 rehearsal (cf. Gardiner et al., 1974). The latter assumption was supported by the finding of moderately positive associations between WMC and retrieval accuracy that were similar in magnitude in each of those tasks. However, we acknowledge the limitation that our inferences are based on indirect evidence, as we did not directly measure List 1 rehearsal during the interpolated task.

In contrast to previous studies, we did not observe interpolated-retrieval effects on list isolation in summary scores of correct recall or intrusions as measures of accessibility and monitoring (cf. Divis & Benjamin, 2014; Jang & Huber, 2008; Sahakyan & Hendricks, 2012). Although these findings were inconsistent with our theory-driven hypotheses, interpolated LTM retrieval has not always led to greater list isolation than has interpolated STM retrieval. For example, Jang and

Huber (2008) found list isolation differences in correct recall between interpolated semantic generation and two-back tasks using a variant of the Shiffrin (1970) list-before-last paradigm. In contrast, Pastötter et al. (2011) did not find any differences in correct recall when comparing variants of the same tasks in a multiple-list learning paradigm. Pastötter et al. did, however, find lower prior-list intrusions when subjects completed an interpolated semantic generation as compared to a two-back task, suggesting that there was at least some effect of the interpolated-task manipulation, possibly on the monitoring of intrusions. The mixed findings from prior studies suggest that task details influence whether interpolated LTM retrieval will increase list isolation. The present study adds to these findings by showing that the use of category exemplar generation and two-back tasks in a multitrial dual-list free recall paradigm was insufficient to show interpolated-retrieval effects on list isolation in summary score measures.

On the basis of the summary score measures reported here, one could conclude that the rates of context change did not differ between the interpolated-task conditions. However, the results from more fine-grained analyses of correct recall probabilities distributed across input positions suggest, provisionally, that interpolated LTM accelerated context change between lists relative to interpolated STM retrieval, but that this acceleration depended on individual differences in WMC. Following studies that had examined differences in free recall dynamics across varying levels of WMC (e.g., Spillers & Unsworth, 2011; Unsworth, 2007, 2009, 2016; Unsworth & Engle, 2007), our theoretically guided exploratory analyses found that List 1 primacy in SP curves was greater following interpolated category exemplar generation than following two-back completion, but only for higher-WMC subjects. We also found an analogous pattern in FRP functions. From these results, we concluded that higher-WMC subjects experienced more rapid contextual change when retrieving from LTM in the interpolated task. This individual-differences conclusion is consistent with existing views of the relationship between WMC and free recall (e.g., Delaney & Sahakyan, 2007; Sahakyan et al., 2014; Unsworth & Engle, 2007). We consider the implications of these results for models of free recall and WMC next.

Models of free recall and WMC

Temporal context model. The temporal context model of episodic memory (e.g., Howard & Kahana, 2002; Lohnas et al., 2015; Polyn et al., 2009) proposes that contextual information becomes associated with items during encoding, and that retrieval success is determined by the extent to which the test context matches the study context and the extent to which study list context can be reinstated. Importantly, in the context of the present study, the model assumes that each item cues retrieval of associated preexperimental context, which serves to update the previous state of the list context. The model further assumes that retrieval from LTM is required to trigger this contextual updating. Thus, the model predicts that context updating should be more rapid when interpolated retrieval is from LTM rather than STM. Finally, the model proposes that STM does not play a role in serial position or contiguity effects, which leads to the prediction of an absence of primacy effects in correct recall.

Inconsistent with model predictions, neither the correct recall nor intratrial intrusion summary scores obtained here showed evidence for differences in list accessibility between the interpolated-task conditions. Furthermore, although free recall functions conditionalized on WMC provided provisional evidence for more rapid context change following interpolated LTM

retrieval for higher-WMC subjects in List 1 primacy effects, the temporal context model would not have predicted these differences, nor would it have predicted primacy effects more generally. One major limitation of this model, from our perspective, is that its predictions are tested by modeling group-level data, which may obscure underlying subject-by-condition interactions (but see Healey & Kahana, 2014). A fruitful direction for future descendants of this model would be to consider how individual differences in context processing, which can partly be measured by WMC, predict differences in context drift and reinstatement.

The dual-store buffer model. As a descendant of Atkinson and Shiffrin (1968), the buffer model proposed by Lehman and Malmberg (2009, 2013) claims the existence of a limited-capacity rehearsal buffer that serves two key functions related to context processing: rehearsal and compartmentalization. *Rehearsal* maintains items in consciousness in order to facilitate context-to-item associations, whereas *compartmentalization* allows subjects to actively drop items from the buffer in order to more effectively process new items. The inclusion of a buffer process in this model creates more flexibility for predictions concerning individual differences in context processing than does the temporal context model. We interpret the buffer model as predicting that higher-WMC subjects, who have superior rehearsal and compartmentalization processes, should form stronger context-to-item associations and better update contextual representations when retrieving from LTM relative to lower-WMC subjects. The buffer model also assumes that control processes serve to monitor output decisions. Therefore, we also interpret the model as predicting that higher-WMC subjects should reject proportionally more intrusions than lower-WMC subjects.

The present results from correct recall production conditionalized on input position, and examined separately for groups varying in WMC, are largely consistent with the predictions from the buffer model. Specifically, the finding that the most pronounced advantage in List 1 primacy (for SP and FRP curves) for the two-back over the generation condition was for higher-WMC subjects suggests that those subjects most effectively processed the context change induced by interpolated LTM retrieval. The buffer model would explain this difference as reflecting higher-WMC subjects being best equipped to drop items from the buffer during interpolated LTM retrieval and to maintain more items during STM retrieval. This combination of abilities would lead to greater differences in the contextual states associated with each interpolated task. However, the results showing greater response initiation from early list positions in the List 2 condition (an immediate test) for higher- than for lower-WMC subjects were inconsistent with the buffer model. The buffer model predicts that subjects with less temporary storage capacity should initiate retrieval from early list positions more often than subjects with more capacity, as those with less capacity should compensate by retrieving from early list position first. Thus, this model may require modification to accommodate individual differences in retrieval initiation strategies that vary with WMC. Finally, the present results showing moderate positive associations between WMC and intrusion monitoring align with predictions of the buffer model.

WMC theory. Although it is not a direct descendant of the Atkinson and Shiffrin (1968) model, the WMC theory from Unsworth, Engle, and colleagues (e.g., Unsworth & Engle, 2007) also includes an active control process that can be engaged across various tasks. WMC theory is similar to the buffer model, in that subjects are assumed to vary in the amount of information

they can actively maintain in working memory. However, according to WMC theory, the main functions of WMC are not limited to maintenance, because WMC also acts in the service of retrieval and monitoring processes. According to the WMC account, maintenance serves to sustain the activity of new information, especially in the face of distraction and interference. In addition, retrieval serves to reinstate context cues to discriminate target memories from distractors. Finally, monitoring serves to evaluate the source of retrieved context.

Assuming that interpolated LTM retrieval accelerates internal context change, WMC theory predicts that higher-WMC subjects should experience the greatest difference in context change between the interpolated-retrieval conditions. That is, higher-WMC subjects should be most likely to sustain their attention to task features, which would heighten the differences between interpolated-retrieval conditions. This would allow higher-WMC subjects to more effectively retrieve preexperimental context during an interpolated LTM retrieval task and to maintain the current context when completing an interpolated STM task. As with the buffer model, this would lead to a greater difference in context updating between interpolated tasks. Finally, the improved monitoring of higher-WMC subjects should lead to more accurate evaluations of the veracity of their retrievals.

Overall, the present results aligned reasonably well with the predictions of WMC theory, which predicts many outcomes in recall performance similar to those from the buffer model. WMC theory also makes explicit predictions about monitoring accuracy that aligned well with the present results, showing moderate positive associations between WMC and response monitoring for both correct recalls and intrusions in many of the trial-type and interpolated-task combinations. WMC theory also correctly predicted a broader distribution of response initiation across input positions in the List 2 condition for higher-WMC subjects, whereas the buffer model predicted the opposite. However, WMC theory did not explicitly predict the WMC-based differences in FRP functions that we observed in the List 1 condition (delayed free recall) showing overall greater response initiation from primacy positions for higher-WMC subjects. This finding points to the role of strategic differences in retrieval initiation that should be further explored to inform future model development. Collectively, these findings indicate that WMC theory and the buffer model can account for the majority of the individual differences observed here, but WMC theory had a slight advantage in predicting the outcomes.

Concluding remarks

In the present experiment, we examined the interactive effects of interpolated LTM retrieval and WMC on list isolation in dual-list free recall. The results did not support the prediction derived from retrieved-context models that interpolated LTM retrieval should lead to greater list isolation effects in summary scores aggregated across subjects, relative to interpolated STM retrieval. However, exploratory analyses of correct recall conditionalized on both input position and WMC provided provisionary support for predictions from buffer and WMC models that context change resulting from interpolated LTM retrieval would be most pronounced for higher-WMC subjects. Importantly, these models could accommodate the observed effects, because they include a control process of the sort originally proposed by Atkinson and Shiffrin (1968, 1971) that could explain individual-level effects. A comprehensive context-based account of interpolated retrieval

effects will therefore require consideration of individual-level differences in active control processes that govern the contents of consciousness.

Author note

The research was supported by internal funding from the University of North Carolina at Greensboro awarded to C.N.W. For their assistance with collecting and coding data, we thank Alexis Blackwell, Marina Hutcherson, Caroline Infante Arismendi, Cayla Kitts, Carson Peske, and Anna Warner.

References

Aslan, A., Zellner, M., & Bäuml, H. (2010). Working memory capacity predicts listwise directed forgetting in adults and children. *Memory*, *18*, 442–450.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposal system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–195). New York, NY: Academic Press.

Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American*, *225*, 82–91.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York, NY: Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bousfield, W. A., & Rosner, S. R. (1970). Free vs. uninhibited recall. *Psychonomic Science*, *20*, 75–76.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33A*, 497–505. https://doi.org/10.1080/14640748108400805

Delaney, P. F., & Sahakyan, L. (2007). Unexpected costs of high working memory capacity following directed forgetting and context change manipulations. *Memory & Cognition*, *35*, 1074–1082.

DeLosh, E. L., & McDaniel, M. A. (1996). The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1136–1146. https://doi.org/10.1037/0278-7393.22.5.1136

Divis, K., & Benjamin, A. (2014). Retrieval speeds context fluctuation: Why semantic generation enhances later learning but hinders prior learning. *Memory & Cognition*, *42*, 1049–1062.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review*, *62*, 145–154. https://doi.org/10.1037/h0048509

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage.

Gardiner, J. M., Thompson, C. P., & Maskarinec, A. S. (1974). Negative recency in initial free recall. *Journal of Experimental Psychology*, *103*, 71–78.

Healey, M. K., & Kahana, M. J. (2014). Is memory search governed by universal principles or idiosyncratic strategies? *Journal of Experimental Psychology: General*, *143*, 575–596. https://doi.org/10.1037/a0033715

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269–299. https://doi.org/10.1006/jmps.2001.1388

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Jang, Y., & Huber, D. E. (2008). Context retrieval and context change in free recall: Recalling from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 112–127. https://doi.org/10.1037/0278-7393.34.1.112

Kahana, M. J., Dolan, E. D., Sauder, C. L., & Wingfield, A. (2005). Intrusions in episodic recall: Age differences in editing of overt responses. *Journal of Gerontology*, *60B*, P92–P97. https://doi.org/10.1093/geronb/60.2.P92

Kahana, M. J., Howard, M. W., Zaromb, F., & Wingfield, A. (2002). Age dissociates recency and lag recency effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 530–540. https://doi.org/10.1037/0278-7393.28.3.530

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General*, *130*, 169–183. https://doi.org/10.1037/0096-3445.130.2.169

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, *132*, 47–70. https://doi.org/10.1037/0096-3445.132.1.47

Keppel, G., Postman, L., & Zavortink, B. (1967). Response availability in free and modified free recall for two transfer paradigms. *Journal of Verbal Learning and Verbal Behavior*, *6*, 654–660.

Klein, K. A., Shiffrin, R. M., & Criss, A. H. (2007). Putting context in context. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger III* (pp. 171–189). New York, NY: Psychology Press.

Lehman, M., & Malmberg, K. J. (2009). A global theory of remembering and forgetting from multiple lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 970–988. https://doi.org/10.1037/a0015728

Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review*, *120*, 155–189. https://doi.org/10.1037/a0030851

Lohnas, L. J., Polyn, S.M., & Kahana, M. J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review*, *122*, 337–363.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*, 203–208. https://doi.org/10.3758/BF03204766

Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 322–336. https://doi.org/10.1037/0278-7393.31.2.322

Meier, M. E., & Kane, M. J. (2013). Working memory capacity and Stroop interference: Global versus local indices of executive control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 748–759.

Meier, M. E., Smeekens, B. A., Siliva, P. J., Kwapil, T. R., & Kane, M. J. (2018). Working memory capacity and the antisaccade task: A microanalytic–macroanalytic investigation of individual differences in goal activation and maintenance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 68–84.

Mensink, G.-J., & Raaijmakers, J. G. W. (1988). A model for interference and forgetting. *Psychological Review*, *95*, 434–455. https://doi.org/10.1037/0033-295X.95.4.434

Miller, A. L., & Unsworth, N. (2018). Individual differences in working memory capacity and search efficiency. *Memory & Cognition*, *46*, 1149–1163. https://doi.org/10.3758/s13421-018-0827-3

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. New York, NY: Cambridge University Press.

Morey, R. D., & Rouder, J. N. (2018). BayesFactor: Computation of Bayes factors for common designs (R package version 0.9.12-4.2). Retrieved from https://CRAN.R-project.org/package=BayesFactor

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K.-H. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 287–297. https://doi.org/10.1037/a0021801

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*, 129–156. https://doi.org/10.1037/a0014420

R Development Core Team. (2008). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation*(Vol. 14, pp. 207–262). New York, NY: Academic Press. https://doi.org/10.1016/S0079-7421(08)60162-0

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, *28*, 164–171.

Roediger, H. L., III, & Payne, D. G. (1985). Recall criterion does not affect recall level or hypermnesia: A puzzle for generate/recognize theories. *Memory & Cognition*, *13*, 1–7. https://doi.org/10.3758/BF03198437

Rohrer, D. (1996). On the relative and absolute strength of a memory trace. *Memory & Cognition*, *24*, 188–201.

Rohrer, D., & Wixted, J. T. (1994). An analysis of latency and interresponse time in free recall. *Memory & Cognition*, *22*, 511–524. https://doi.org/10.3758/BF03198390

Sahakyan, L., Abushanab, B., Smith, J. R., & Gray, K. J. (2014). Individual differences in contextual storage: Evidence from the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 873–881.

Sahakyan, L., & Hendricks, H. E. (2012). Context change and retrieval difficulty in the list-before-last paradigm. *Memory & Cognition*, *40*, 844–860.

Sahakyan, L., & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1064–1072. https://doi.org/10.1037/0278-7393.28.6.1064

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize*? Journal of Research in Personality*, *47*, 609–612.

Shiffrin, R. M. (1970). Forgetting, trace erosion or retrieval failure? *Science*, *168*, 1601–1603. https://doi.org/10.1126/science.168.3939.1601

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145–166. https://doi.org/10.3758/BF03209391

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *Dialogue*, *26*, 4–7.

Soriano, M. F., & Bajo, M. T. (2007). Working memory resources and interference in directed forgetting. *Psicologia*, *28*, 63–85.

Spillers, G. J., & Unsworth, N. (2011). Variation in working memory capacity and temporal–contextual retrieval from episodic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1532–1539. https://doi.org/10.1037/a0024852

Unsworth, N. (2007). Individual differences in working memory capacity and episodic retrieval: Examining the dynamics of delayed and continuous distractor free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 1020–1034. https://doi.org/10.1037/0278-7393.33.6.1020

Unsworth, N. (2009). Variation in working memory capacity, fluid intelligence, and episodic recall: A latent variable examination of differences in the dynamics of free recall. *Memory & Cognition*, *37*, 837–849. https://doi.org/10.3758/MC.37.6.837

Unsworth, N. (2016). Working memory capacity and recall from long-term memory: Examining the influences of encoding strategies, study time allocation, search efficiency, and monitoring abilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 50–61.

Unsworth, N., & Brewer, G. (2010). Variation in working memory capacity and intrusions: Differences in generation or editing? *European Journal of Cognitive Psychology*, *22*, 990–1000.

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2011). Inter- and intraindividual variation in immediate free recall: An examination of serial position functions and recall initiation strategies. *Memory*, *19*, 67–82. https://doi.org/10.1080/09658211.2010.535658

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2013). Focusing the search: Proactive and retroactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1742–1756.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*, 104–132. https://doi.org/10.1037/0033-295X.114.1.104

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*, 498–505. https://doi.org/10.3758/BF03192720

Unsworth, N., & Spillers, G. J. (2010). Variation in working memory capacity and episodic recall: The contributions of strategic encoding and contextual retrieval. *Psychonomic Bulletin & Review*, *17*, 200–205. https://doi.org/10.3758/PBR.17.2.200

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2012). Evidence for noisy contextual search: Examining the dynamics of list-before-last recall. *Memory*, *20*, 1–13.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*, 289–335. https://doi.org/10.1016/j.jml.2003.10.003

Wahlheim, C. N., Ball, H., & Richmond, L. L. (2017). Adult age differences in production and monitoring in dual-list free recall. *Psychology and Aging*, *32*, 338–353.

Wahlheim, C. N., & Huff, M. J. (2015). Age differences in the focus of retrieval: Evidence from dual-list free recall. *Psychology and Aging*, *30*, 768–780.

Wahlheim, C. N., Richmond, L. L., Huff, M. J., & Dobbins, I. G. (2016). Characterizing adult age differences in the initiation and organization of retrieval: A further investigation of retrieval dynamics in dual-list free recall. *Psychology and Aging*, *31*, 786–797.

Ward, G., & Tan, L. (2004). The effect of the length of to-be-remembered lists and intervening lists on free recall: A reexamination using overt rehearsal. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1196–1210. https://doi.org/10.1037/0278-7393.30.6.1196

Appendix 1

Table 3. Category exemplar generation task materials

Practice		**Block 1**				**Block 2**				**Block 3**				**Block 4**			
Category: Animal		**Category: Color**		**Category: Fruit**		**Category: Sport**		**Category: Weather**		**Category: Bird**		**Category: Insect**		**Category: Fish**		**Category: Instrument**	
Exemplar	**Fragment**	**Exemplar**	**Fragment**	**Exemplar**	**Fragment**	**Exemplar**	**Fragment**	**Exemplar**	**Fragment**	**Exemplar**	**Fragment**	**Exemplar**	**Fragment**	**Exemplar**	**Fragment**	**Exemplar**	**Fragment**
bear	b _ a _	aqua	aq _ _	apple	a _ pl _	baseball	b _ s _ ba _ _	blizzard	bl _ zzar _	canary	c _ n _ ry	ant	a _ t	beta	be _ a	bass	b _ _ s
elephant	e _ _ ph _ _ _	black	bl _ _ k	banana	b _ n _ _ a	basketball	b _ sk _ _ _ a_ l	cloud	c _ ou _	cardinal	c _ rd _ na _	bee	b _ e	carp	c _ r _	cello	c _ llo
giraffe	g _ _ af _ e	blue	bl _ _	blueberry	bl _ _ b _ _ r _	bowling	b _ wli _ _	earthquake	e _ rt _ qu _ ke	crow	cro _	beetle	b _ et _ e	catfish	c _ tfi _ _	clarinet	cl _ ri _ et
goat	g _ _ t	brown	br _ _ _	cherry	ch _ rr _	golf	g _ l _	flood	f _ oo _	dove	dov _	butterfly	bu_ te _ _ ly	cod	c _ d	drum	d _ _ m
horse	h _ r _ _	green	g _ _ e _	grape	g _ _ p _	hockey	h _ ck _ y	hail	h _ il	eagle	ea _ l _	centipede	c _ nti _ ed _	dolphin	d _ lp _ _ n	flute	fl _ _ e
lion	li _ n	indigo	ind _ _ o	lemon	l _ mo _	rugby	r _ gb _	hurricane	h _ rric _ _ _	falcon	f _ l _ on	cricket	c _ ick _ t	flounder	fl _ und _ r	guitar	g _ i _ ar
lizard	l _ _ ar _	magenta	m _ g _ _ ta	lime	l _ me	running	r _ nni _ _	rain	r _ i _	finch	f _ nch	flea	fl _ a	guppy	gu _ py	harmonica	h _ rm _ _ ica
moose	m _ os _	maroon	m _ roo _	mango	m _ ng _	skiing	sk _ i _ g	sleet	sl _ _ t	flamingo	f _ ami _ _ o	fly	f _ y	herring	h _ rri _ g	harp	h _ rp
rabbit	r _ bb _ _	orange	or_ _ ge	peach	p _ ac _	soccer	s _ cc _ _	snow	sn _ _	hawk	h _ w _	gnat	g _ at	minnow	m _ nn _ w	keyboard	ke _ bo _ r _
raccoon	r _ _ co _ _	pink	p _ _ k	pear	pe _ _	softball	s _ ft _ all	storm	st _ _ m	owl	ow _	ladybug	l _ dyb _ _	salmon	s _ lm _ n	organ	o _ ga _
squirrel	sq _ _ _ re _	purple	p _ rp_ _	pineapple	p _ nea _ ple	swimming	sw _ mm _ _ _	thunder	t _ und _ _	parrot	p _ r _ ot	moth	mot _	shark	s _ _ rk	piano	pi _ _ o
tiger	t _ ge _	red	r _ _	plum	p _ u _	tennis	t _ nn _ _	tornado	t _ rna _ _	raven	r _ _ en	roach	roac _	snapper	snap _ er	saxophone	s _ xop _ _ ne
turtle	t _ _ t _ e	teal	t _ a _	raspberry	r _ _ pb _ r _ y	track	t _ _ ck	tsunami	ts _ n _ mi	robin	r _ bi _	spider	s _ ide _	trout	tr _ ut	trumpet	trum _ _ t
wolf	w _ _ f	white	w _ i _ _	strawberry	s _ _ awb _ _ _ _	volleyball	vo_ l _ yb _ _ _	typhoon	typ _ o _ n	seagull	s _ ag _ ll	wasp	w _ s _	tuna	tu _ a	tuba	t _ b _
zebra	z _ _ r	yellow	y _ ll _	watermelon	wa _ er _ e l _	wrestling	wr _ st _ _ ng	wind	w _ _ d	sparrow	sp _ rr _ w	worm	wor _	whale	wh _ l _	violin	v _ ol _ n

Note: Exemplars from each category are displayed in alphabetical order, which is different from the order in which they appeared in the experiment

Appendix 2

Table 4. Two-back task materials

Practice		**Block 1**				**Block 2**				**Block 3**				**Block 4**			
Item	**Trial Type**	**Item**	**Trial Type**	**Item**	**Trial Type**	**Item**	**Trial Type**	**Item**	**Trial Type**	**Item**	**Trial Type**	**Item**	**Trial Type**	**Item**	**Trial Type**	**Item**	**Trial Type**
q	2-back (p1)	w	Foil (p1)	b	2-back (p1)	r	2-back (p1)	x	2-back (p1)	h	Foil (p1)	w	2-back (p1)	n	2-back (p1)	l	2-back (p1)
w	2-back (p1)	w	Foil (p2)	m	Single	c	2-back (p1)	h	Single	h	Foil (p2)	k	Single	s	Single	t	Single
q	2-back (p2)	j	Single	b	2-back (p2)	r	2-back (p2)	x	2-back (p2)	s	Single	w	2-back (p2)	n	2-back (p2)	l	2-back (p2)
w	2-back (p2)	l	Foil (p1)	n	Foil (p1)	c	2-back (p2)	w	Single	v	2-back (p1)	l	2-back (p1)	g	Single	n	Foil (p1)
t	2-back (p1)	l	Foil (p2)	n	Foil (p2)	n	2-back (p1)	j	2-back (p1)	p	Single	t	2-back (p1)	k	2-back (p1)	n	Foil (p2)
t	Foil (p2)	n	2-back (p1)	h	Foil (p1)	f	Single	q	Single	v	2-back (p2)	l	2-back (p2)	l	2-back (p1)	k	2-back (p1)
t	2-back (p2)	x	Single	h	Foil (p2)	n	2-back (p2)	j	2-back (p2)	c	2-back (p1)	t	2-back (p2)	k	2-back (p2)	r	Single
b	2-back (p1)	n	2-back (p2)	t	Single	h	Foil (p1)	p	2-back (p1)	q	2-back (p1)	b	Foil (p1)	l	2-back (p2)	k	2-back (p2)
v	Single	b	2-back (p1)	p	2-back (p1)	h	Foil (p2)	f	2-back (p1)	c	2-back (p2)	b	Foil (p2)	z	Foil (p1)	v	Single
b	2-back (p2)	p	2-back (p1)	k	Single	q	2-back (p1)	p	2-back (p2)	q	2-back (p2)	v	2-back (p1)	z	Foil (p2)	z	Foil (p1)
r	Foil (p1)	b	2-back (p2)	p	2-back (p2)	m	Single	f	2-back (p2)	t	Foil (p1)	g	Single	v	2-back (p1)	z	Foil (p2)
r	Foil (p2)	p	2-back (p2)	w	2-back (p1)	q	2-back (p2)	d	Foil (p1)	t	Foil (p2)	v	2-back (p2)	t	Single	h	2-back (p1)
g	2-back (p1)	h	2-back (p1)	d	2-back (p1)	v	Foil (p1)	d	Foil (p2)	z	2-back (p1)	c	Foil (p1)	v	2-back (p2)	p	2-back (p1)
x	Single	z	Single	w	2-back (p2)	v	Foil (p2)	r	Foil (p1)	g	Single	c	Foil (p2)	q	Foil (p1)	h	2-back (p2)
g	2-back (p2)	h	2-back (p2)	d	2-back (p2)	b	Single	r	Foil (p2)	z	2-back (p2)	x	Single	q	Foil (p2)	p	2-back (p2)

The trial type abbreviations are as follows: Single = character that only appeared once; Foil (p1) = the first presentation of a one-back foil repetition; Foil (p2) = the second presentation of a one-back foil repetition; 2-back (p1) = the first presentation of a two-back repetition; 2-back (p2) = the second presentation of a two-back repetition